FACT-BASED VISUAL QUESTION ANSWERING USING KNOWLEDGE GRAPH EMBEDDINGS

BY

KIRAN RAMNATH

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois Urbana-Champaign, 2021

Urbana, Illinois

Adviser:

Professor Mark Hasegawa-Johnson

Abstract

Humans have a remarkable capability to learn new concepts, process them in relation to their existing mental models of the world, and seamlessly leverage their knowledge and experiences while reasoning about the outside world perceived through vision and language. Fact-based Visual Question Answering (FVQA), a challenging variant of VQA, requires a QA-system to mimic this human ability. It must include facts from a diverse knowledge graph (KG) in its reasoning process to produce an answer. Large KGs, especially common-sense KGs, are known to be incomplete, i.e., not all non-existent facts are always incorrect. Therefore, being able to reason over incomplete KGs for QA is a critical requirement in real-world applications that has not been addressed extensively in the literature. We develop a novel QA architecture that allows us to reason over incomplete KGs, something current FVQA state-of-the-art (SOTA) approaches lack due to their critical reliance on fact retrieval. We use KG embeddings, a technique widely used for KG completion, for the downstream task of FVQA. We also present a new image representation technique we call *image-as-knowledge* which posits that an image is a collection of knowledge concepts describing each entity present in it. We also show that KG embeddings hold complementary information to word embeddings. A combination of both metrics permits performance comparable to SOTA methods in the standard answer retrieval task, and significantly better (26% absolute) in the proposed missing-edge reasoning task.

The next research problem pursued is extending the accessibility of such systems to users through a speech interface and providing support to multiple languages, which have not been addressed in prior studies. We present a new task and a synthetically generated dataset to do Fact-based Visual Spoken-Question Answering (FVSQA). FVSQA is based on the FVQA dataset, with the difference being that the question is spoken rather than typed. Three sub-tasks are proposed: (1) speech-to-text based, (2) end-to-end, without speech-to-text as an intermediate component, and (3) cross-lingual, in which the question is spoken in a language

different from that in which the KG is recorded. The end-to-end and cross-lingual tasks are the first to require world knowledge from a multi-relational KG as a differentiable layer in an end-to-end spoken language understanding task, hence the proposed reference implementation is called Worldly-Wise (WoW). WoW is shown to perform end-to-end cross-lingual FVSQA at the same levels of accuracy across three languages - English, Hindi, and Turkish.

To my parents, friends, and family, for their unending love and constant support.

Acknowledgments

The work behind this thesis has led me through a beautiful and humbling process of self-discovery. I would like to acknowledge my privilege that allows me to engage in such fulfilling academic pursuits in times of great troubles around the world.

I owe a debt of gratitude to the many people who were by my side. I thank my adviser Dr. Mark Hasegawa-Johnson, who so graciously took me under his tutelage, gave me the freedom and courage to explore scientific ideas without fear of failure, and helped me grow in the field of AI research. This thesis would not have taken shape without his endless patience and wisdom, his timely and insightful feedback on conducting and communicating research, and most of all, his calming presence that inspires as much as it comforts. I would also like to thank Prof. Chang Yoo from KAIST who provided a great opportunity for collaboration and exploration.

Throughout this journey, I was fortunate to have the support of constant companions as well as new friends and collaborators. A few of them include: Akanksha Agarwal, Suraj Ramnath, Prarthana Ranganathan, Sharanya Subramaniam, Sakshi Srivastava, Siddharth Muralidaran, Aditya Shankar Narayanan, Pulkit Katdare, John Harvill-Bowman, Mahir Morshed, Leda Sari, and Sudheer Salana.

Lastly, but most importantly, none of this would be possible without my parents' love and selfless nourishment that helped me grow to be the individual I am today.

Table of Contents

Chapter 1

Introduction

Question answering has long been considered as an important milestone for achieving artificial general intelligence (AGI), first proposed by Alan Turing as a proxy for judging machine intelligence. In parallel, multi-modal AI tasks have focused on mimicking the way humans interact visuo-linguistically with the outside word. These tasks require an AI-system to reason about information coming from different input sources such as vision, language, audio, etc., to solve a given problem.

The Visual Question Answering (VQA) [1] benchmark is a distillation of these pursuits – it requires a system to answer a question in relation to an image. VQA is challenging because it requires many capabilities such as object detection, scene recognition, and activity recognition in addition to language understanding and commonsense reasoning. Its applications are widespread – information retrieval, personal assistants, online shopping, etc. Crucially however, [1] noted that most of the questions in previous VQA datasets did not require commonsense knowledge residing outside of the image. Humans have a remarkable ability to blend in knowledge from their own prior experiences when answering a question about an image. They therefore introduced the FVQA benchmark, aiming to provide a more challenging set of questions which ensure that the answer to a given question requires some form of external knowledge, not present in the image or the question text. They provide this external information in the form of knowledge graphs (KG), which are multi-relational graphs, storing relational representations between entities. The task of FVQA boils down to retrieving the correct entity as the answer most relevant to an image. For example, in Fig. 1.1 the fact triple – (Cat, IsA, Mammal) is the external information required to answer the question; with the correct answer being 'Cat'. Many different KGs cover different kinds of relationship types to cater to domain-specific knowledge.

In the context of question answering, it is worth asking - What does it mean to reason like a human? While there can be no single way to approach this, we identify two main capabilities that such systems must display - i) reasoning

Question - Which entity in this image is a mammal?
Supporting fact - [[A cat]] is [[a mammal]]
Subject, Predicate, Object - (Cat, IsA, Mammal)
Answer Source - Image
Answer - Cat

Figure 1.1: Example of a fact-based visual question

over incomplete yet deducible knowledge for QA, and ii) providing interpretable symbolic reasoning over concepts to explain its deductive process. Inspired by how humans leverage prior experiences while interacting in a new language, we pursue two additional capabilities: i) answering questions directly based on audio inputs, ii) transferring symbolic concepts across languages for QA in under-resourced languages. The rest of this work describes our research attempts at imparting these capabilities.

1.1 FVQA over incomplete KGs with KG embeddings

The deployment of such a question-answering application to real-world systems must contend with practical issues. For example, large production-scale knowledge graphs are compiled by parsing large amounts of text from the web, thereby suffering from incompleteness. All previous methods assumed the completeness of the accompanying KG for their working. In a real-world application, however, it is reasonable to expect users to ask questions that the graph does not yet contain. Therefore, it is important to build systems that can reason about facts that are not yet present, but are true, or can somehow be inferred to be true. The primary

contribution of our work [2] is a method Seeing is Knowing (SiK) that permits FVQA to reason about commonsense facts that are absent from the knowledge graph using KG embeddings. When used with a more general score function we define, the resultant framework works well in both complete and incomplete KG settings. KG embeddings permit us to offer two additional contributions to the previous state-of-the-art (SOTA) in FVQA: an *image-as-knowledge* representation of visual information, and a co-attention method for combining visual and textual inputs. Image-as-knowledge represents the image as the span of the KG embedding vectors for the entities found in it. Representing an image with textual semantics has been attempted before [3], but not using KG embeddings; KG embeddings provide robustness to incomplete KGs by encoding information about the graph structure. Finally, SiK when used in a standalone fashion over incomplete KGs is more time-efficient during inference. It is $O(m)$ as it only needs to reason over existent nodes in the network. To the best of our knowledge, our work is the first one to apply KG embeddings to a VQA task.

1.2 End-to-end cross-lingual spoken question answering over KGs

Now imagine being able to ask your voice assistant a question not just in English but in any language, to learn some trivia about your favorite movie-star. Previous methods in FVQA and other tasks have mostly focused on well-resourced languages [4, 1]. These languages generally also have mature automatic speech recognition (ASR) systems and language models. The accompanying knowledge graphs also tend to be limited to languages that are well-resourced [5, 6, 7]. Against this background, it is worthwhile to think of building end-to-end systems which directly use speech signals as input, that can readily harness huge knowledge repositories stored in another language, instead of requiring *tabula rasa* learning.

This work [8] formulates a natural extension to the FVQA task called Fact-based Visual Spoken-Question Answering. This requires a system to perform cross-lingual spoken question answering over images and KGs. We show that the versatility of neuro-symbolic KG embeddings allows for a seamless transfer of knowledge across different languages. The aim of this work is to motivate research in the direction of cross-lingual speech tasks that attempt to leverage common-sense repositories from one language for use with another.

Chapter 2

Background

2.1 Knowledge graphs

Simmons [9] first proposed the idea of using multi-relational graph structures as semantic networks to encode semantic information. Various projects such as [10, 5, 11, 12, 13] have since approached the task of representing the way entities or concepts are related to each other. These KGs are stored as collections of fact-triplets of $f = (\texttt{subject}, \texttt{predicate}, \texttt{object})$. KGs are usually directed graphs, and a fact triplet denotes the type of directed edge connecting its subject with its object. Informally, a fact is also said to comprise of $f = (\texttt{head}, \texttt{relationship}, \texttt{tail})$. Halford et al. [14] and Murphy [15] have found that such linked representations studied herein resemble the cognitive processes adopted during human reasoning as well.

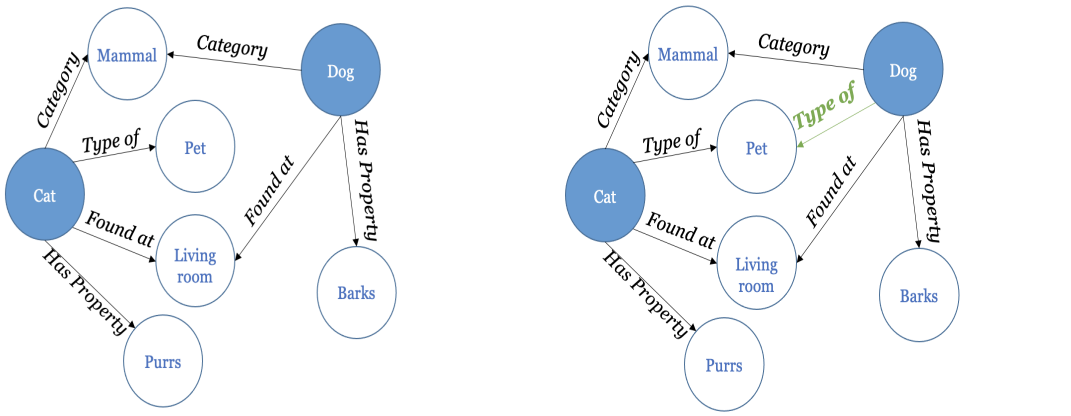
KGs can be said to fall into two broad categories: open-world or closed-world. Closed-world KGs necessitate that facts not present in the KG are false. On the other hand, open-world KGs relax this requirement; therefore, a fact not present currently in the KG could simply be missing instead of being false. Formally, the observed knowledge graph is denoted by $\mathcal{G} \subset \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, where $\mathcal{E}$ is the set of all entities, $\mathcal{R}$ is the set of all relationship-types. Open-world KGs are therefore characterized as $\mathcal{G} \subset \mathcal{G}_T \subset \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, where $\mathcal{G}_T$ is the unobserved set of all true facts in the world that the KG seeks to represent.

2.2 KG embeddings

Research paradigms seeking to reason on the basis of incomplete knowledge graphs include the tasks of predicting missing links, disambiguating duplicate entries, and clustering entities based on similar attributes or connections. An important notion

is that 'an entity is known by the neighbors it keeps', thus the semantic structures imposed by edge-constraints can encode useful information that can be leveraged for several downstream semantic tasks.

Knowledge graph embeddings attempt to encode these graph structures by embedding the entities and relationships in a high-dimensional space. The dimensionalities of both the entity and relation embeddings N_e and N_r respectively are usually set to be equal. The motivation to learn KG embeddings [16, 17, 18, 19, 13, 20] is to predict whether a given (h, r, t) triple is true or false. Formally, a score function ϕ learns node-features through the mapping $\phi(h, r, t) : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \rightarrow [0, 1]$, where $h, t \in \mathcal{E}$ are the head and tail entities, and $r \in \mathcal{R}$ is the relation connecting the two. The embeddings (h, r, t) are learned so that the score $\phi(.)$ is high for facts not just in $\mathcal{G}$, but also in $\mathcal{G}_T$, and low for facts outside $\mathcal{G}_T$. Figure 2.1 shows a dummy example of how KG embeddings can predict missing links in the KG.



a) Incomplete KG b) Predicting absent facts using KG embeddings.

Figure 2.1: Example of KG completion using KG embeddings. Embeddings learn that an entity which is a mammal and also found in the living room should also be a type of pet.

Distance-based models learn embeddings h, r and t in order to minimize the distance between t and $f(h, r)$, for some projection function $f(\cdot)$. Two such distance-based models widely studied in the literature are TransE [16] and RotatE [17]. TransE models $f(h, r) = h + r$. RotatE models $f(h, r)$ as a Hadamard product, $h \circ r$. For RotatE, the embedding vectors h, r and t belong to the complex plane. During training, r is constrained to have unit-norm, so that the element-wise multiplication with r is a rotation of the complex-vector h. Another suitable model we study for our KG completion task is the Entity-Relation Multi-Layer Perceptron (ERMLP) [13]. It uses an MLP to produce the score $\phi(h, r, t)$ for each fact triple.

2.3 KGQA

Knowledge-graph question answering (KGQA) is the task of answering questions based on the facts in a KG. 'Simplequestions' was one of the first textual KGQA benchmarks [21]. Lukovnikov et al. [22] attempted this by using character-level embeddings on both the questions as well KG entities. Some works have since tried to approach KGQA using KG embeddings - works such as KEQA [23] and CFO [24] use translational embeddings as entity and relationship representations and retrieve the correct entity that minimizes an appropriate distance metric, achieving SOTA results on SimpleQuestions, FB2M, and FB5M. EmbedKGQA [25] also used embedding-based reasoning over incomplete KGs on the Webquestions [26] and MetaQA [27]. These were language-only benchmarks, but this approach has not yet been tested in multimodal reasoning tasks. Different from their work, our task involves the visual modality, as well as reasoning over a common-sense KG.

Among KGQA baselines including the visual modality, the OKVQA benchmark [28] provides outside common-sense knowledge in the form of supporting text. KVQA [29] provides KGs as outside knowledge, and the existing baselines focused on face-recognition and entity-linking to answer several different types of questions. Both baselines incorporate knowledge from the KG using a neural network parse of the fact text, not KG embeddings.

2.4 Visual Question Answering

In the last section, we discussed the task of performing textual question answering which incorporate facts from an external KG. The other crucial piece to Fact-based Visual Question Answering (FVQA) is the incorporation of the visual modality in such a setting. Thus, before discussing FVQA, it is important to shed some light on the broad area of Visual Question Answering (VQA) - tracing its beginnings and discussing some of its variants that have been pursued. DAQUAR [32] was the very first benchmark that merged the research paradigm of question answering to include the visual modality, aiming to introduce a visual analog to the Turing test. VQA 2.0 [4] is one of the more famous and comprehensive benchmarks in this area, which improved upon the previous iteration [33]. It was guided by the finding that models attempting the previous benchmark were simply overfit to the text-modality. A system was able to answer questions well even if the

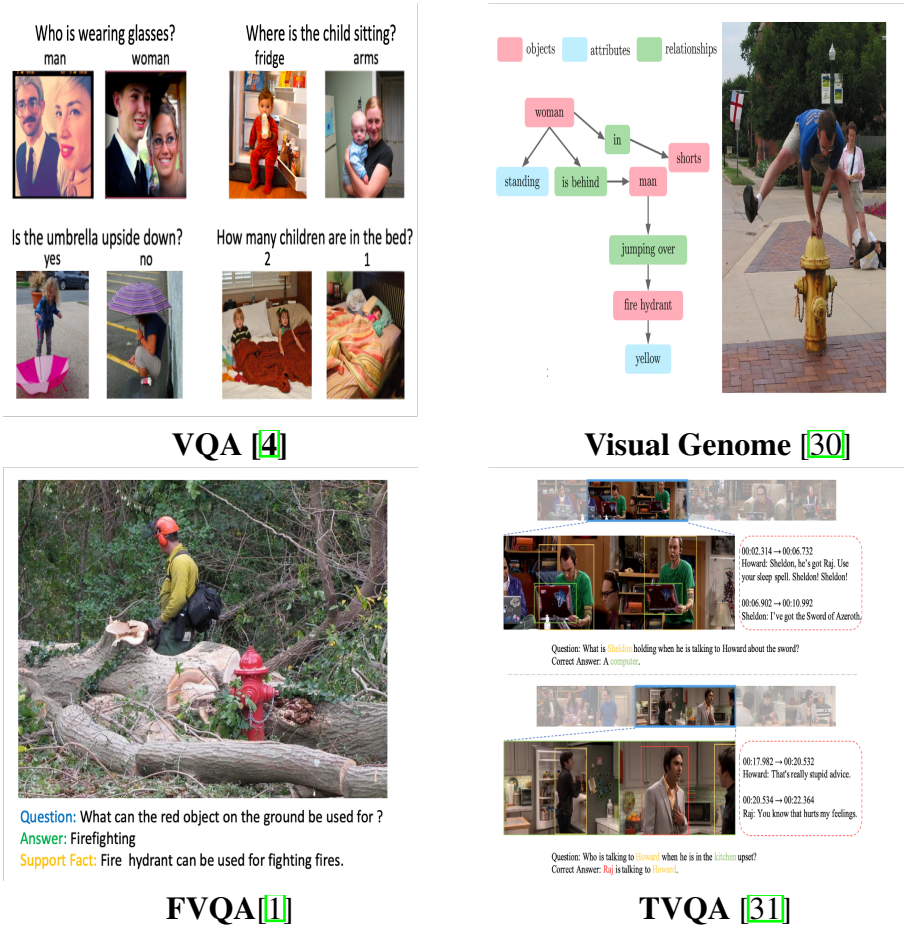


Figure 2.2: Some examples of VQA benchmarks

input image was removed. This happened because for every given question, the answers had very low diversity, and a simple question-to-answer choice mapping could be learned without paying attention to the image. VQA 2.0 [4] presented an enhanced dataset which had confounding images for every question thereby increasing entropy of the conditional distribution $P(Ans|Question)$, such that a degenerate system providing the same answer to the same question must necessarily be false for at least one image in the dataset.

Since then, several challenging variants of VQA [34, 35] have been proposed that incorporate one or more kinds of reasoning capabilities as described in Figure 2.2. A few of the important directions pursued in VQA research include: a) VQA based on dense spatial understanding [30, 36], b) common-sense knowledge [1, 29, 28], and most recently c) video-based question answering [31, 37].

The anatomy of a general VQA system has three broad components – visual processing module, language processing module, and finally a co-processing module which fuses the two representations and leads to generating the correct answer (see Fig. 2.3). The correct answer could either be one of many classes, short phrases, or whole sentences.

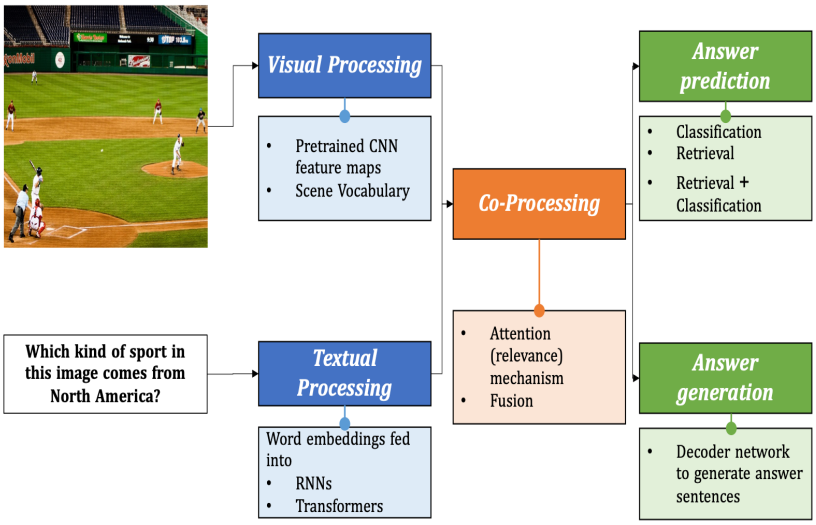


Figure 2.3: Anatomy of a VQA system

2.5 Fact-based VQA

The Fact-based Visual Question Answering (FVQA [1]) benchmark was designed using questions that obey the following condition: for each (question,image,answer) triplet in the dataset ($(q_i, I_i, y_i) \in \mathcal{D}$), there exists exactly one supporting fact in the knowledge graph ($f_j = (h, r, t) \in \mathcal{G}$) such that the correct answer y_i is either the head or the tail of f_j, and such that at least one of the two entities is visible in the image.

Different from other single-hop KGQA datasets, answers for FVQA can occur either as head or tail of a fact, instead of only as head. In context of FVQA, this

translates to there being two types of answers to any question. The final answer could be an entity found in the image - this entity is called the key visual concept (KVC). Or it could be some knowledge about this entity that is to be retrieved from the knowledge graph; such answers are denoted as knowledge base (KB). Nonetheless, both types of questions require reasoning over facts from the KG.

The accompanying KG is also diverse, comprising facts from three individual KGs: Webchild [6], ConceptNet [7], and DBPedia [5]. The DBPedia KG mainly covers hypernym relationships - i.e. denotes which category an entity belongs to. The ConceptNet project is a more general counterpart of WordNet [38], in that it conveys common-sense facts relating not just atomic entities but also by incorporating long phrases of words to improve the expressive power of the KG. This however, adds to more sparsity of such entities. Finally, Webchild provides many different kinds of comparative relationships. These comparative relations are considered as a single edge-type for the task of FVQA. In total, the dataset contains 13 relations: $\mathcal{R} \in \{$CATEGORY, HASPROPERTY, RELATEDTO, ATLOCATION, ISA, HASA, CAPABLEOF, USEDFOR, DESIRES, PARTOF, RECEIVESACTION, CREATEDBY, COMPARATIVE$\}$. Table 2.1 provides examples of a fact for each type of relationship in the KG. The dataset consists of 2190 images sampled from the ILSVRC [39] and the MSCOCO [40] datasets. The accompanying KG consists of roughly 194500 facts concerning 88606 entities. Roughly 82% of the questions have a key visual concept (KVC) as the answer, whereas 18% have information from the knowledge base (KB) as the answer.

Table 2.1: Number of facts available in the FVQA KG for each type of relationship. [1]

Type of relationship	**# facts**	**Sample fact**
CreatedBy	96	{photo, CreatedBy, camera}
ReceivesAction	344	{piano, ReceivesAction, play}
PartOf	762	{wheel, PartOf, car}
HasA	1665	{train, HasA, carriage}
HasProperty	2813	{airplane, HasProperty, noisy}
Desires	3358	{cat, Desires, be feed}
CapableOf	5837	{fish, CapableOf, swim}
IsA	6011	{chicken, IsA, meat}
AtLocation	13,683	{bus, AtLocation, bus stop}
Category	35,152	{cat, Category, mammal}
Comparative	38,576	{ship, Comparative, plane}
RelatedTo	79,789	{ithaca, RelatedTo, island}

Answering questions in FVQA is to solve for

$$\hat{y} = \underset{e \in \mathcal{E}}{\operatorname{argmax}}\ p(y = e \mid q, I, \mathcal{G}), \tag{2.1}$$

i.e., finding the most probable entity as the answer given a question q and image I, and given the graph $\mathcal{G}$. Our formulation of the missing edge reasoning task considers $\mathcal{G}_T$ instead of $\mathcal{G}$.

Previous approaches which relied on complete accompanying KGs try to decompose this as:

$$\hat{y} = \underset{e \in f^*}{\operatorname{argmax}}\ p(e \mid f^*) : \quad f^* = \underset{f \in \mathcal{G}}{\operatorname{argmax}}\ p(f \mid q, I). \tag{2.2}$$

Wang et al. [1] attempted FVQA as a parsing and querying problem, constructing 32 different templates of graph-queries, and classifying each image-question pair as requiring one of these templates. Executing the chosen query returns candidate facts, and simple keyword matching techniques further prune the retrieved facts. Their next work approached VQA as reading comprehension [3]. It extends the previous work, where it represents the image as textualized knowledge, which can then be fused seamlessly with the question modality. However, the downstream architecture still remains the same, which first chooses one of 32 pre-defined graph query templates to retrieve the correct fact first. Straight to the facts (STTF) [41] approached FVQA by directly learning to retrieve supporting facts using a deep neural network, where each fact-entity was represented using lexical semantic representations. This approach, however, does not make use of KG structure.

Out-of-the box (OOB) [42], the previous SOTA, then extended this approach by using local neighborhood-based reasoning via a graph convolutional network (GCN) [43] to answer each question. OOB answers the query by constructing a subgraph based on the query and image, then applying GCN reasoning to the subgraph; the subgraph construction stage has an inference time complexity of $O(n\ log\ n)$. Our proposed method is also extended using a similar subgraph construction to enhance its performance. However, the standalone method without the subgraph construction still performs with comparable accuracy in $O(m)$ time. The GCN models local interactions through message passing between nodes, but potentially ignores global structures that could be useful for this task.

Chapter 3

FVQA over Incomplete Knowledge Graphs

3.1 A new image representation

In the previous chapter, we saw that KG embeddings learn useful features that help complete the KG by encoding the existing structures of the KG. Inspired by such concise semantic concept-representations, we ask a natural question (as illustrated in Fig. 3.1) - Is it possible to represent an image as a collection of such knowledge concepts? This is especially attractive in the context of the FVQA task, since answering a question is to retrieve an entity from the KG. Such an approach would therefore allow the system to represent the image in the same space as that in which the answer retrieval happens. Besides the mathematical and implementational characteristics of such a representation, this advocates more broadly that a visual sensory input serves as a bridge to the conceptual domain, the latter of which has been built through a sophisticated accumulation of prior experiences. The following sections first delineate at a high-level how such an image representation will be used for question-answering. Thereafter, we describe each of the required building blocks in order to enable such a representation.

3.2 Our approach - Seeing is Knowing

Our architecture follows a rank-and-retrieval approach to produce the correct answer. It processes the input representations of the image and question in order to predict a best-guess answer vector in the same space as the concept embeddings. Since the correct answer to each question could be either the subject (KVC) or object (KB) of some fact from the KG, our architecture is motivated to use two identical but separately trained architectures to produce two different query vectors. The proposed architecture for FVQA is shown in Fig. 3.2. As shown, a given image

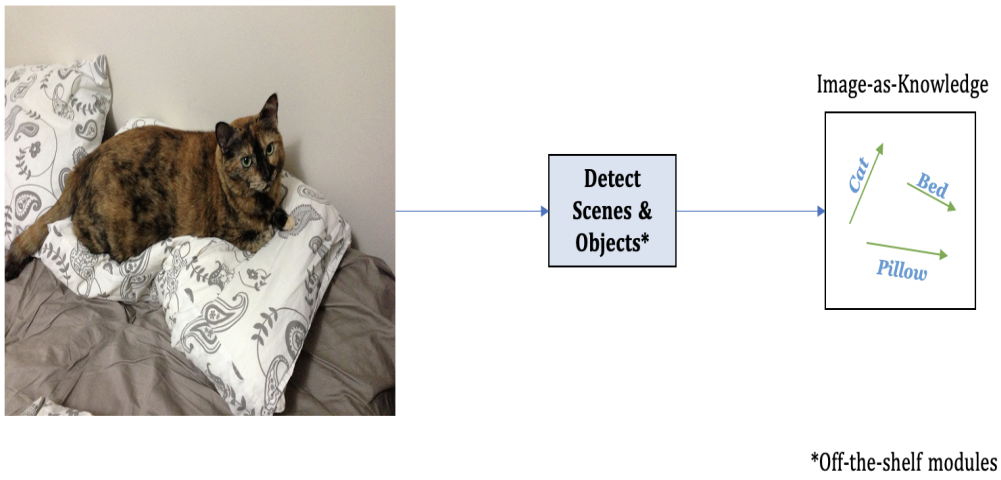


Figure 3.1: Representing an image as knowledge

I and query q are combined via co-attention to form the two entity query vectors, $f_{KVC}(q, I)$ and $f_{KB}(q, I)$. The KG is then queried for the answer to the question, according to

$$\hat{y}(q|I) = \begin{cases} \underset{e \in \mathcal{E}}{\operatorname{argmax}} f_{KVC}(q, I)^T e & g_{KVC}(q) = 1 \\ \underset{e \in \mathcal{E}}{\operatorname{argmax}} f_{KB}(q, I)^T e & g_{KVC}(q) = 0, \end{cases} \tag{3.1}$$

where the gating function $g_{KVC}(q) \in \{0, 1\}$ is equal to 1 if the text of the question indicates that the answer is visible in the image and equal to 0 otherwise.

The rest of this section addresses representations of the entities, image, and query, the information fusion functions, the gating function, and the loss function.

3.2.1 KG representation

We train KG embeddings by setting up a self-supervised learning task. A surrogate binary classification problem is designed which assigns truth-probabilities to every fact triplet, such that the probability is high for edges which must be true and low for edges which must be false. In the process, high-dimensional embeddings are learned for entities and relationships that help the score-function assign these truth-probabilities to all facts in the network. To learn a non-trivial and useful binary classifier, we must provide the model with negative examples (examples of edges that do not exist in the knowledge graph AND are false), lest it simply learn to inflate the model-parameters and classify all facts as true. However, in the case

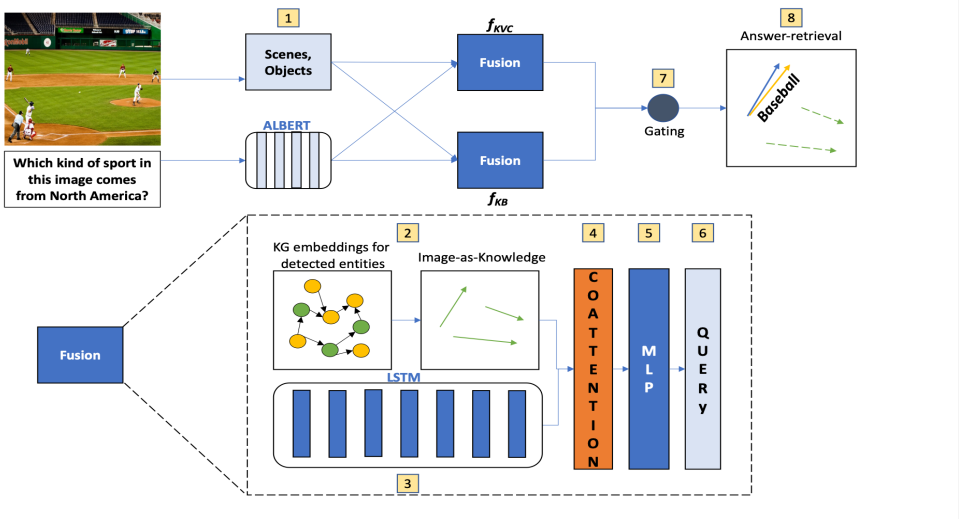


Figure 3.2: Seeing is Knowing architecture. (1) Scenes, objects, & actions are detected in the image. (2) For detected entities, we retrieve their KG embedding representations. The span of these embedding vectors represents the image-as-knowledge. (3) Lexical semantic vectors for each word in the query are accumulated via an LSTM. (4) The joint image-question encoding is derived using a co-attention mechanism described in Fig. 3.3, then (5) passed through a multi-layer perceptron, whose (6) last layer is used as a pair of queries that are (7) gated, and (8) used to retrieve the entity to answer the question.

of open-world KG embeddings, this poses a unique chicken-and-egg problem - KG embeddings are supposed to solve the very problem of predicting which missing edges are true, yet they need some false edges to actually learn a good classifier. Sampling negatives for KG embeddings therefore has to be done with care.

Some heuristics have been empirically found to work well in overcoming this problem. Under the locally closed world assumption (LCWA) [13], negative samples can be generated by randomly corrupting the tail entity of existing facts, i.e.

$$\texttt{if } (h, r, t) \in \mathcal{G} \texttt{ then } (h, r, t') \in \mathcal{G}_T^c \; \forall t' \in \mathcal{E}. \tag{3.2}$$

It was found by [44] that it is often too easy to classify such negative samples as false, and therefore they proposed an architecture based on generative adversarial networks (GAN) [45] to construct such negative examples. However it is well-known that training a separate GAN is often difficult in the face of unstable training regimes, mode collapse [46], etc.

In our work, to sample hard negatives to train KG embeddings, we use a self-adversarial negative sampling strategy proposed by [17] which attempts to leverage the best of both approaches - i.e. ease of generating negative samples as well as

generating meaningful negative samples. First, the adversarial examples $f'_j = (h_i, r_i, t'_i)$ are generated by randomly corrupting t'_i from each observed edge $f_i = (h_i, r_i, t_i)$. Next, these samples are weighted by their truth-probability as estimated by the learned embeddings h and t, and by the corresponding score function $\phi(h, r, t)$. This avoids the learning of another generator model, and instead proposes using the network's own probability parameterization to assign weights to negative samples.

The KG embedding loss function penalizes the network when a true edge has a low truth-probability, and a false edge has a high truth-probability. But some false facts may be more difficult for the model to classify than the others. Sun et al. [17] found that the loss function should reflect this, which is why each false fact's contribution to the loss is scaled by the truth-probability assigned by the network during training. Thus, false edges with a higher truth-probability are penalized more heavily than false edges with lower truth-probabilities. A total of n adversarial examples are generated for each true fact, and used to train discriminative embeddings using noise contrastive estimation [47]. Thus the knowledge graph embedding loss $\mathcal{L}_{KGE}$ includes the negative log probability that each observed edge is true ($\ln \sigma(\phi(f_i))$), and the expected log probability that the adversarial edges are false ($\ln \sigma(-\phi(f'_j)) = \ln\left(1 - \sigma(\phi(f'_j))\right)$):

$$\mathcal{L}_{KGE} = -\sum_{i=1}^{|\mathcal{G}|} \left(\ln \sigma\big(\phi(f_i)\big) + \sum_{j=1}^{n} p_i(f'_j) \left[\ln \sigma\big(-\phi(f'_j)\big)\right] \right), \tag{3.3}$$

where the probability $p_i(f'_j)$ is the softmax-probability of a false fact f'_j, generated from a given ground truth fact f_i and computed within the n negative samples generated. The logits are scaled using α (a temperature hyperparameter) as:

$$p_i(f'_j) = \frac{\exp(\alpha\, \phi(f'_j))}{\sum_{k=1}^{m} \exp(\alpha\, \phi(f'_k))}. \tag{3.4}$$

Eq. (3.3) is used to train embeddings of the head (h) and tail (t), which are applied to the FVQA task as described in the next several subsections. Eq. (3.3) also trains relation embeddings (r) and MLP weights for the ERMLP scoring function (w_{MLP}); these quantities are not used for the downstream FVQA task.

Our method for generating corrupt triples for each fact is compliant with the LCWA, and different from [17] which also generates negatives as corrupt head entities for a (r, t) pair.

3.2.2 Language representation

For representing the words in the question, we use the last layer of contextual ALBERT [48] embeddings without finetuning for FVQA training. After passing through an LSTM, we use the hidden state representation for each word w_t for further processing as $q_i^t = h(w_t)$ as the question representation.

To obtain a single condensed vector representation for the question, we calculate attention weights for each word in the question, and subsequently a self-attention-weighted encoding for the question as:

$$A(q_i) = \sum_{t=1}^{|q_i|} \alpha_q^t \, q_i^t, \quad \alpha_q^t = \frac{\exp(w_{\alpha_q}^T q_i^t)}{\sum_{t=1}^{|q_i|} \exp(w_{\alpha_q}^T q_i^t)}, \tag{3.5}$$

where α_q^t, w_{α_q} are respectively the attention paid to word w_t, and the weight vector from which it is computed.

3.2.3 Image representation

STTF [41] found that providing raw feature maps from a pre-trained convolutional network like ResNet [49] or VGG [50] actually hurt performance on FVQA. It instead advocated using a symbolic image representation scheme, which in their case was a simple multi-hot vector (i.e. multiple one-hot variables) over all concepts classes. The image-as-knowledge representation is similarly motivated, but instead represents each image as a collection of high-dimensional knowledge vectors.

Objects: We use Torchvision's Coco object-detector, a faster RCNN detector [51] with ResNet50 backbone [49], and feature pyramid network [52], which detects 80 object classes. We also use a detector [53] trained on an OpenImages 600-class detection task. We include classes present in ImageNet 200 plus those in [54] to maximize overlap with the dataset used with FVQA.

Scenes: We use a wideresnet [55] detector trained on MIT365 places dataset [56] and consider the 205 classes that were used in constructing the FVQA KG.

Overall, we detect 540 visual concepts. Having detected visual concepts in each image, we represent the i^{th} image as a collection of entities, $I_i = [e_i^1, \ldots, e_i^m] \in \mathbb{R}^{\mathrm{N_e \times m}}$, where N_e is the embedding dimension, and m is the number of visual concepts detected in the image. We detect a maximum of $m = 14$ visual concepts

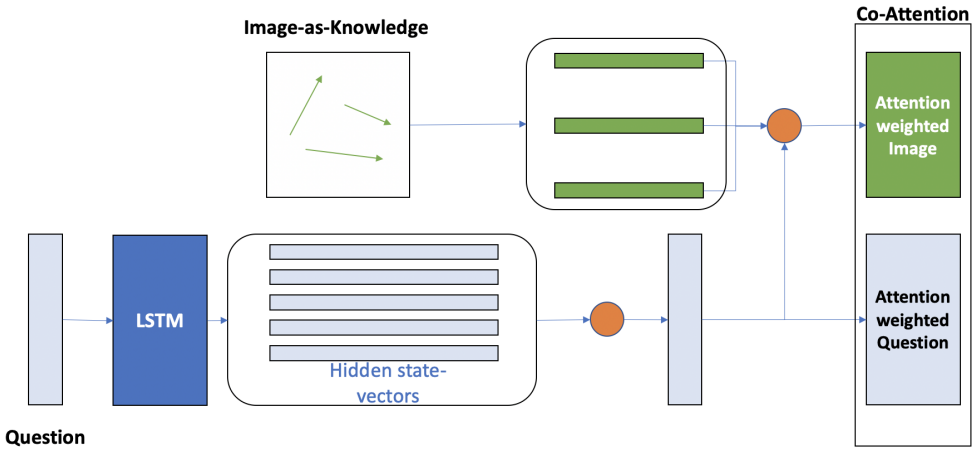


Figure 3.3: Image and query are fused using the co-attention mechanism depicted here. First, self-attention computes a weighted summary of the query (bottom orange circle). Second, the query is used to compute an attention-weighted summary of the concepts in the image (top orange circle). The resulting image query is a vector drawn from the span of the entities present in the image.

in each image. Our findings imply that for FVQA reasoning, an image is best represented as a collection of KG entity embedding vectors e_i^j, apparently because these KG embeddings encode the graph structure and background information necessary to be able to answer questions. We think this is an important finding, and we show its effect on the answer prediction accuracy.

Having detected the presence of visual concepts in each image and a contextually weighted sum of the question representation, we now describe the image-as-knowledge representation, where the image is represented as a linear combination of the knowledge vectors of its constituent visual concepts, thus is a vector drawn from the span of the knowledge vector for concepts detected in it.

Using $A(q_i)$ as a query, we compute the attention-weighted summary of the image:

$$A(I_i) = \sum_{j=1}^{m} \alpha_I^j \, e_i^j, \quad \alpha_I^j = \frac{\exp\left(w_{\alpha_I}^T \begin{bmatrix} A(q_i) \\ e_i^j \end{bmatrix}\right)}{\sum_{k=1}^{m} \exp\left(w_{\alpha_I}^T \begin{bmatrix} A(q_i) \\ e_i^k \end{bmatrix}\right)}, \tag{3.6}$$

where $\alpha_I^j, w_{\alpha_I}, e_i^j$ are respectively the attention paid to concept j in the image, the weight vector from which it is computed, and the j^{th} concept present in the image.

Both $A(I_i)$ and $A(q_i)$ learn a mapping: $R^{N_e \times m} \rightarrow R^{N_e}$, which is the attention-based weighted average of its inputs

3.2.4 Fusion functions f_{KVC} and f_{KB}

These attention-weighted image and question encodings compute joint image-question encodings via late fusion as:

$$f_{KVC}(q_i, I_i) = h\left(A(I_i), A(q_i); w_{KVC}\right) \tag{3.7}$$

$$f_{KB}(q_i, I_i) = h\left(A(I_i), A(q_i); w_{KB}\right), \tag{3.8}$$

where $h(\cdot)$ is a two-layer fully connected network with ReLU activation functions. Using $f_{KVC}(q_i, I_i) = A(I_i)$, i.e., the attention weighted image encodings, to directly retrieve the answer significantly reduces accuracy, suggesting the need for successive fully connected layers to add capacity.

3.2.5 Gating function g_{KVC}

The fusion functions f_{KVC} and f_{KB} are trained to retrieve two different entities from the KG, either of which might be the answer to the question. The gating function, g_{KVC}, selects one of these two. The gating function is a sigmoid fully connected layer applied to the final output state of an LSTM, whose input is the query, and which is trained using binary cross entropy so that $g_{KVC} = 1$ if the correct answer is a key visual concept in the image. This gating function is similar to those used by FVQA and STTF [1, 41], except that during retrieval, the gating function is quantized and used to select an entity query before retrieving the entity from the KG, rather than after retrieving the fact.

3.2.6 Loss function

A summary of all parameters is provided in Table 3.1. Entity embeddings are learned in order to minimize the loss function in Eq. (3.3), then all parameters are jointly trained in order to minimize the cosine distance between the ground truth entity, y_i, and the network output:

$$\mathcal{L}_{FVQA} = \frac{1}{n}\sum_{i=1}^{n}\left(1 - y_i^T \hat{y}(q_i|I_i)\right), \tag{3.9}$$

where $\hat{y}(q_i|I_i)$ is as given in Eq. (3.1).

3.2.7 Text-augmented composite score

Guided by the observation that KG embeddings often encode KG semantics but not lexical (word-token) or distributional (word embedding) semantics (see Fig. 4.1), we further define a text-augmented score to enhance the answer retrieval accuracy. This helps the answer retrieval to take advantage of overlapping words between the question and the fact entities. Firstly, a pruning strategy as introduced by OOB [42] is used (see Alg. 1) to retrieve the top 100 most relevant facts. The words in the question and the fact are passed through a tokenizer and stop-word remover (denoted by `PROCESS()` in the algorithm). Next, using GLoVE 100-D embeddings [57], a similarity score $\eta(f_k)$ is computed for each KG fact f_k, by first computing the highest cosine similarity $S(w_t)$ for each word w_t in the fact against every word in the question and detected visual entities. The top 80% scoring words from each fact are retained, and these similarity scores are averaged to assign a score to each fact for a given question. The top 100 scoring facts are retrieved for each question, denoted as $\mathcal{F}_{100}$. Answer retrieval then takes place from entities within this reduced set of facts $\mathcal{F}_{100}$. We define the score function, a convex combination of three complementary metrics, as follows:

$$\hat{y}(q|I) = \underset{e \in \mathcal{F}_{100}}{\operatorname{argmax}} \left\{ \lambda_1 K(q,I)^T e + \lambda_2 J(q,e) \; + \lambda_3 D(q,e) \right\}$$
$$\text{s.t.} \sum_{k=1}^{3} \lambda_k = 1.$$

Here, $K(q,I)^T e$ is the KG similarity score of an entity with the query produced by the SiK network. $J(q,e) = \max_{f \in \mathcal{F}_{100}, e \in f} j(q,f)$, where $j(q,f)$ is the Jaccard (word-token) similarity between the question and one of the retrieved facts f. An entity may be a part of more than one fact in $\mathcal{F}_{100}$; $e \in f$ means that e is either the head or tail of f. $D(q,e) = \max_{f \in \mathcal{F}_{100}, e \in f} d(q,f)$ where $d(q,f)$ is the similarity between its fact and a question based on averaged GloVe embeddings.

Algorithm 1: Pruning entity search space to set of most relevant facts $\mathcal{F}_{100}$

for $q_i, I_i \in \mathcal{D}$ **do**
  Obtain token set $\mathcal{T} = \texttt{PROCESS}(q_i, I_i)$;
  for $f_k \in \mathcal{G}$ **do**
    for $w_t \in \textit{PROCESS}(f_k)$ **do**
      $S(w_t) = \max_{c \in \mathcal{T}} \frac{g(w_t)^T g(c)}{|g(w_t)||g(c)|}$;
    end
    $\mathcal{M} = \{w_j : \underset{j \in |f_k|}{\texttt{argsort}}\ S(w_j)\}$;
    $\eta(f_k) = \sum_{j=1}^{0.8*|\mathcal{M}|} S(w_j)$;
  end
  $\mathcal{F}_{100} = \{f_k : \underset{k \in |\mathcal{G}|}{\texttt{argsort}}\ \eta(f_k)\}_{k=1}^{100}$;
end

3.3 Experimental setup

3.3.1 KG embeddings training

For training KG embeddings for the standard FVQA task, the entire KG is split into 80% training set and 20% test set. The embedding dimensions for both entity and relation embeddings is $N_e = N_r = 300$. The batch size used is 1000. All the KG embeddings are trained for 25,000 epochs. Adam optimizer is used for which the learning rate was initialized as 0.01 and then it is scaled down by a factor of 0.1 after every 10,000 epochs. The hyper-parameter search for the learning rate was performed by choosing among values in the set $\{0.0001, 0.001, 0.01, 0.1\}$. The temperature hyper-parameter α for the self-adversarial probability parameterization is set to 1 for all experiments. The number of adversarial samples n generated for each positive sample is 16.

ERMLP is parameterized as a three-layer neural network. The size of the first layer is $3N_e$ since it takes the head, rel, and tail embeddings as input. Subsequent layers are $2N_e$ and N_e in size, which are finally capped by a single sigmoid unit to output the truth-probability $\phi(h, r, t)$. The activation functions used by the hidden layers are the rectified linear unit (ReLU), which outputs $max\{0, x\}$ for an input x. All networks are fully connected and none of the networks uses the dropout layer.

Accuracy for KG embeddings is measured using the standard metrics: Hits @1, Hits @3, Hits @10. These determine how often each correct tail/head gets

Table 3.1: Summary of all trainable parameters. Head (h) and tail (t) use identical embedding vectors.

Parameters	Used for	Loss Function
h, r, t, w_{MLP}	KG embeddings	$\mathcal{L}_{KGE}$
$w_{\alpha_q}, w_{\alpha_I}, w_{KVC}$, w_{KB}, θ_{LSTM}	Answer-retrieval based on image and question	$\mathcal{L}_{FVQA}$
w_g, θ_{LSTM}	Choose answer source $\in \{f_{KVC}, f_{KB}\}$	Cross entropy

ranked in the top 1, 3, or 10 ranked facts for each ground-truth (h, r)/(r, t) pair. Mean rank is a metric often used to gauge the performance of KG embeddings. It measures the mean rank of each true fact in the dataset when ranked by its truth-probability for a given (h, t) pair. An allied metric is mean reciprocal rank = $\frac{1}{|\mathcal{D}|}\Sigma_i \frac{1}{R_i}$. While we report these metrics, we do note that these metrics are not best suited for common-sense KGs for reasons mentioned before. However a more refined analysis of these methods is left for future work.

3.3.2 FVQA training

We report Hits @1 and Hits @3 for each algorithm. All numbers are based on averaging results across five train - test splits provided with the dataset. The number of questions varies slightly, but roughly half fall into the training set, and half into the testing set, for each split. Stochastic gradient descent with a batch size of 64 trains f_{KVC} and f_{KB} for 250 steps with a learning rate of 0.01, reduced by 0.1 every 100 epochs, and a weight decay of 1e-3. Fully connected layers use a dropout probability of 0.3. The gating function g_{KVC} is trained for 20 steps with step-size of 0.1.

GPUs provided by Google Colab are used to train all models. Our heaviest KG embedding technique (ERMLP) takes roughly 3 hours to train, while one train split for f_{KVC} takes roughly 30 minutes. Subgraph construction for each question takes about 1s, and 250 epochs to train our implementation of OOB took roughly 3 hours.

3.3.3 Baselines

Results in the standard QA task for the models FVQA, STTF, and OOB are taken from the respective papers since the codes for these systems have not been made available. HieCoAtt denotes using a hierarchical co-attention network [58] as implemented in [1]. AvgEmbed denotes a method which compares STTF's and OOB's fact representation methodology with our architecture to correctly perform missing-edge reasoning. We use averaged 300-dimensional GLoVE embeddings [57] to represent each entity. We also report results from our OOB implementation in the incomplete KG setting.

Chapter 4

Results and Discussion

The gating function achieves an accuracy of 96%, the same as the accuracy reported in STTF [41]. While we do not use fact-relationship predictions for retrieving the answer in our architecture, we find that using our architecture $f(q_i, I_i)$ by setting $I_i = 0$ and using the ground-truth relationship embedding as supervision instead of the entity, gives us an average accuracy of 74%, again close to the levels reported by [41].

We discuss below the performance of other modules of our architecture.

4.1 Ablation-study for FVQA accuracy

4.1.1 $\mathcal{F}_{100}$ fact recall and impact on FVQA

We only observe 63% fact recall within $\mathcal{F}_{100}$ across the entire dataset as opposed to 84.8% as reported by [42]. This recall drops further to 53% if filtering for the top three relationships as OOB does in its best-performing model. This difference is prominent, and FVQA accuracies reported by OOB at such levels of fact recall are significantly lower. One factor for lower recall is noisy visual detections, whose keywords impact fact retrieval. Therefore, with access to similar visual detections as the other methods, SiK could potentially achieve even higher accuracy. Furthermore, owing to the lower fact retrieval recall, our OOB implementation could not replicate the reported performance in the standard QA task. This also highlights the crucial reliance of OOB on correct fact retrieval and further substantiates our observation of its low performance when the KG is incomplete.

Table 4.1: FVQA accuracy (IaK - image-as-knowledge, Adv - self-adversarial negative sampling, NR - not reported)

Technique	Hits@1	Hits@3
HieCoAtt	33.7 ± 1.18	50.00 ± 0.78
FVQA top-1 qq-mapping	52.56 ± 1.03	59.72 ± 0.82
STTF	$62.2 \pm NR$	$75.6 \pm NR$
OOB top-1 rel	$\mathbf{65.8 \pm NR}$	$\mathbf{77.32 \pm NR}$
OOB top-3 rel	$69.35 \pm NR$	$80.25 \pm NR$
Seeing is Knowing		
AvgEmbed, IaK	27.77 ± 0.72	32.43 ± 0.94
ERMLP – Adv, Multihot	51.51 ± 1.14	64.89 ± 1.37
ERMLP, IaK	53.16 ± 0.79	$63.9 \pm .63$
ERMLP -Adv, IaK	$\mathbf{54.38 \pm 0.94}$	$\mathbf{65.76 \pm 0.5}$
Seeing is Knowing - text augmented		
ERMLP - Adv, IaK ($\lambda_1 = 0.4$, $\lambda_2 = 0.3$,$\lambda_3 = 0.3$)	$\mathbf{60.82 \pm 0.94}$	$\mathbf{77.14 \pm 0.50}$
ERMLP - Adv, IaK ($\lambda_1 = 0$, $\lambda_2 = 0.5$,$\lambda_3 = 0.5$)	39.76 ± 1.08	60.84 ± 0.67

4.1.2 Performance of Seeing is Knowing

Table 4.1 shows the performance of Seeing is Knowing in the standard FVQA task. When the required edge is present in the graph, standalone SiK underperforms OOB by 11%. This is enhanced using the composite score, which sees a 6% point improvement in Hits@1 accuracy, while it matches the previous SOTA for Hits@3 accuracy. Quite remarkably, text-augmented SiK works best when roughly equal coefficients are used for all three similarities.

4.1.3 Incomplete KGs

Table 4.2 demonstrates the robustness of SiK in the incomplete KG setting. Here we discuss two of them - one where only the QA-related facts are missing and another where 50% of the KG is missing. In both cases, we see standalone SiK and text-augmented SiK only lose some of their accuracy. Our implementation of OOB, and AvgEmbed both underperform SiK by over 25%. Setting $\lambda_1 = 0$ in the score also leads to poor performance, highlighting the importance of KG features for FVQA.

The performance of the ERMLP algorithm on the KG completion task is shown in Table 4.3, where we can see that its performance remains robust despite various

Table 4.2: Missing edge experiments

Incomplete KG - Only QA facts missing		
Technique	**Hits@1**	**Hits@3**
AvgEmbed, IaK	27.77 ± 0.72	32.43 ± 0.94
OOB top-3 rel	28.58 ± 0.01	43.56 ± 0.01
ERMLP -Adv, IaK	53.45 ± 0.77	65.1 ± 1.41
ERMLP - Adv, IaK ($\lambda_1 = 0.34$, $\lambda_2 = 0.33$,$\lambda_3 = 0.33$)	$\mathbf{55.13 \pm 0.97}$	$\mathbf{73.04 \pm 0.54}$
ERMLP - Adv, IaK ($\lambda_1 = 0$, $\lambda_2 = 0.5$,$\lambda_3 = 0.5$)	26.8 ± 0.90	49.24 ± 0.62
Incomplete KG - 50% KG occluded		
AvgEmbed, IaK	27.77 ± 0.72	32.43 ± 0.94
OOB top-3 rel	27.53 ± 0.01	41.56 ± 0.01
ERMLP -Adv, IaK	52.29 ± 0.86	62.74 ± 0.69
ERMLP - Adv, IaK ($\lambda_1 = 0.34$, $\lambda_2 = 0.33$,$\lambda_3 = 0.33$)	$\mathbf{55.14 \pm 1.1}$	$\mathbf{72.95 \pm 0.38}$
ERMLP - Adv, IaK ($\lambda_1 = 0$, $\lambda_2 = 0.5$,$\lambda_3 = 0.5$)	26.78 ± 0.856	49.20 ± 0.63

Table 4.3: KG embedding accuracy for ERMLP.

Sampling	**MR**	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**
Adversarial	**11194**	**0.156**	**0.132**	**0.152**	**0.197**
Uniform	14907	0.122	0.09	0.128	0.173
Incomplete KG results					
Incompleteness	**MR**	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**
(QA facts missing)	11071 ± 479	0.162 ± 0.007	0.136 ± 0.003	0.162 ± 0.005	0.204 ± 0.007
(50% KG)	12880	0.144	0.1	0.13	0.161

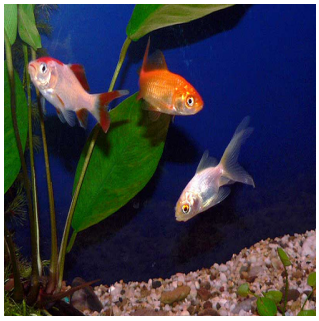
levels of occlusion. More results based on incomplete KGs are discussed in Tables 4.4 and 4.5.

4.1.4 Success and failure cases: The need for text augmentation

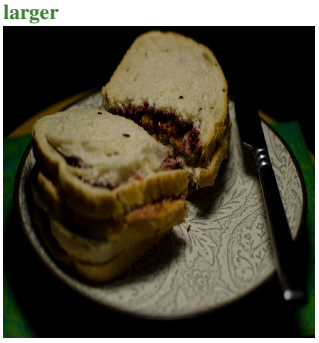
To investigate the only learning module in the score, we look at the failure modes for the standalone SiK architecture. Since the gating function performs with almost perfect accuracy and there are fewer sequential decision-points compared to both STTF and OOB, our main source of error is incorrect entity prediction. Since an answer-prediction is considered correct only if there is an exact match in the entities,

Question 1: What is the difference between the instrument & the violin?
SPO triple: : {Cello, HASPROPERTY, like a violin but larger}
Answer Source: KG
Answer: like a violin but larger
Answer predicted: **like a violin but larger**

Question 2: Which object in this image belongs to the category Herbivorous animals?
SPO triple: {Giraffe, CATEGORY, Herbivorous animals}
Answer Source: Image
Answer: Giraffe
Answer predicted: **Giraffe**



Question 3: What popular pet is in this image?
SPO triple: {Goldfish, ISA, popular pet}
Answer Source: Image
Answer: Goldfish
Answer predicted: **Fish**



Question 4: What object in this image is commonly eaten for lunch?
SPO triple: {Sandwich, ISA, meal commonly eaten for lunch}
Answer Source: Image
Answer: Sandwich
Answer predicted: **Bread**

Question 5: Where can you find the large object in the back of this image?
SPO triple:{Bus, ATLOCATION, Bus stop}
Answer Source: KG
Answer: Bus stop
Answer predicted: **Wait place**

Question 6: What can we find in the place shown in this image?
SPO triple: {Furniture, ATLOCATION, Bedroom}
Answer Source: KG
Answer: Furniture
Answer predicted: **Your House**

Figure 4.1: Success and failure cases of Seeing is Knowing

many inaccuracies stem from cases where the network predicts a semantically valid but different entity. Questions 3-6 in Fig. 4.1 show that similar attributes connected through multi-hop relationships are commonly mistaken by the SiK network. Our model makes fewer mistakes when a KVC is the answer, and they occur when there are other concepts which are highly similar to the ground-truth answer, and can sometimes be a valid answer in their own right. KG embeddings represent entities based on relationships they have with other entities; therefore, it is easy to see why the model could mistake 'sandwich' for 'bread', or 'goldfish' for 'fish'. For the cases when f_{KVC} gives the answer, we see an accuracy of $64.42\% \pm 0.76$, whereas for f_{KB} it is only $4.45\% \pm 1.32$. It must be reiterated that we consider an answer as correct only if there is an exact match between the entity predicted and

the ground-truth. We empirically observe that this stringent but narrow requirement leads to a significant fraction of semantically relevant answers being considered wrong. More failure cases are discussed in Section 4.3.

4.1.5 Image as knowledge

To understand how useful the image-as-knowledge representation is compared to other variants, we compare it to the multihot variant proposed by Narasimhan et al. [42] along with our best performing KG embedding technique. Image-as-knowledge provides a 3-point performance improvement (Table 4.1), apparently because the retrieval of an entity happens in the entity space spanned by the vectors of the IaK representation.

4.1.6 Self-adversarial negative sampling

We see that the accuracy improves for both the downstream FVQA (Table 4.1) and the upstream KG embedding task (Table 4.3) upon introducing self-adversarial negative sampling during KG embedding training. Guu et al. [59] reported that initializing entity vectors as averaged word embeddings yields better performance. We see no significant improvement on doing so compared to randomly initializing the entity vectors.

Table 4.4: KG embedding performance of ERMLP when varying levels of the KG are occluded.

KG Occlusion	**MR**	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**
Only QA facts	11071	0.162	0.136	0.162	0.204
10% of KG	9690	0.175	0.148	0.1779	0.22
20% of KG	10733	0.1613	0.133	0.162	0.213
30% of KG	10437	0.154	0.127	0.156	0.199
40% of KG	11467	0.139	0.116	0.138	0.179
50% of KG	12880	0.144	0.1	0.13	0.161

Table 4.5: FVQA accuracy of Seeing is Knowing using ERMLP when varying levels of the KG are occluded.

KG Occlusion	Hits @1	Hits @3
QA facts	53.45 ± 0.77	65.1 ± 1.41
10% of KG	53.19 ± 0.47	65.67 ± 0.31
20% of KG	53.62 ± 1.01	64.72 ± 0.98
30% of KG	53.15 ± 1.62	64.55 ± 1.23
40% of KG	53.14 ± 0.47	64.77 ± 0.55
50% of KG	52.29 ± 0.86	62.74 ± 0.69

4.2 Experiments with KG occlusion

Table 4.4 shows the performance of ERMLP using standard metrics at different levels of KG incompleteness. As expected, there is a small but monotonous decline in the performance with more and more occlusion. A known shortcoming of these metrics is that when the KG has $(subject, predicate)$ pairs that can lead to more than one $object$ entity, they will consider a different but factually correct entity as incorrect.

Quite remarkably however, there is only a very slight decrease in SiK's performance when using any of these KG embeddings for the downstream FVQA task (see Table 4.5). Only at 50% KG occlusion do we see a statistically significant yet still small drop in performance. This highlights the robustness of using KG embeddings for downstream semantic tasks.

4.3 Qualitative discussion

4.3.1 Success cases

A deeper analysis of the success cases depicts the two main benefits provided by using KG embeddings - i) mitigating visual saliency bias, and ii) robustness to word-rephrasing. Each object's relevance to a given visual question depends solely on its knowledge representation, and not the size of its appearance in the image. For example, in Question 2 in Fig. 4.2, one can see that the guitar is one among many in a room full of objects, but the network is able to reason correctly. Similarly, in Question 8 in Fig. 4.2, the cat is smaller in size, and its pixels overlap with the bicycle. But owing to the image-as-knowledge representation, our network

does not need to contend with partially or fully occluded objects as long as it is prominent enough for an object detector to detect it.

It must be reiterated that the answer retrieval does not happen only within the few objects found inside an image; rather the best aligned entity from the entire KG is chosen, but KG embeddings hold enough discriminative power to enable this retrieval. We can see that the knowledge graph embeddings also successfully convey notions that are worded in a complex manner. For instance, in Question 9 in Fig. 4.2 the complexly worded phrase 'indicates passage of time' is successfully matched with the watch entity that is detected in the image.

4.3.2 Failure cases

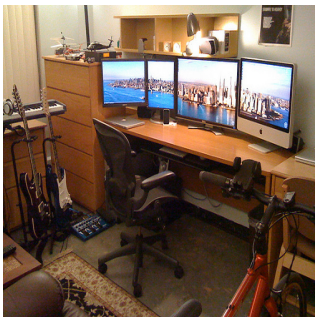
Analyzing the types of errors made by SiK provides interesting insights that further highlight the necessity of text augmentation for answer retrieval. We now discuss some more failure cases by characterizing its inaccuracies. We observe that there are three broad categories of errors:

1. Semantically correct answer, but answer is a different entity
2. Incorrect answer, explainable relevance to question and image
3. Entirely incorrect answers

An empirical study of these errors shows a significant fraction of them fall in the first two categories. Errors in the first category can arise when the model has implicitly pinned down the correct key visual entity and relationship being asked about, but then chose a different entity as the answer (see Questions 4-6 Fig. 4.3). A bed is an object one can lose things in, chairs come in many styles, and a flute can be used for practice. These errors can also arise if the model has chosen to place visual emphasis on an entity that is different from ground truth, but is equally valid given a question (see Question 7 in Fig. 4.3). Here, the model has retrieved the answer 'build a treehouse' that appears in a fact with 'tree', but the ground truth of the question asks about the entity 'lake'.

Errors of the second category can arise in mainly two cases - incorrect language and / or visual grounding. In cases where the visual grounding is off, it could

Question 1: What is this place can be used for?
SPO triple: {Bedroom, UsedFor,Sleeping}
Answer Source: KG
Answer: Sleeping
Answer predicted: **Sleeping**



Question 2: What thing in the image must be tuned frequently?
SPO triple: {guitar, ReceivesAction, tune frequently}
Answer Source: Image
Answer: Guitar
Answer predicted: **Guitar**

Question 3: Which object in this image is used to nail wood?
SPO triple:{hammer, CapableOf, drive nail on wood}
Answer Source: Image
Answer: hammer
Answer predicted: **hammer**

Question 4: What is the object in this image used for?
SPO triple: {IPod, UsedFor, listening to music}
Answer Source: KG
Answer: listening to music
Answer predicted: **listening to music**

Question 5: Which appliance in this image will you use to keep food fresh?
SPO triple: {refrigerator, UsedFor, keep food fresh}
Answer Source: Image
Answer: refrigerator
Answer predicted: **refrigerator**

Question 6: What object in this image is a plant?
SPO triple:{Grass, IsA, Plant }
Answer Source: Image
Answer: Grass
Answer predicted: **Grass**

Question 7: Why the man wears a helmet?
SPO triple:{Helmet, ReceivesAction, wear to protect the head}
Answer Source: KG
Answer: protect the head
Answer predicted: **protect the head**

Question 8: What in this image is capable of kill bird?
SPO triple:{cat, CapableOf, killing bird}
Answer Source: Image
Answer: cat
Answer predicted: **cat**

Question 9: Which object in this image indicates the passage of time?
SPO triple: {Clock, UsedFor, indicate passage of time}
Answer Source: Image
Answer: Clock
Answer predicted: **Clock**

Figure 4.2: Success cases of SiK

Question 1: What is the animal famous for?
SPO triple: {Giraffe, HASPROPERTY, Long necked}
Answer Source: KG
Answer:long necked
Answer predicted: work for day without water

Question 2: Which object in this image is used in spaghetti sauce?
SPO triple: {Spaghetti sauce, RELATEDTO, Tomato}
Answer Source: Image
Answer: Tomato
Answer predicted: Bell Pepper

Question 3: What is the family of the animal in this image?
SPO triple:{Fox, CATEGORY, Canidae}
Answer Source: KG
Answer: Canidae
Answer predicted: Burundian Culture

Question 4: What is the wooden thing used for?
SPO triple: {Flute, USEDFOR, making music}
Answer Source: KG
Answer:making music
Answer predicted: Practice

Question 5: What is the material of the chair?
SPO triple: {Chair, RELATEDTO, Wooden}
Answer Source: KG
Answer: Wooden
Answer predicted: come in many style

Question 6: What is the object in the left of the image used for?
SPO triple:{Bed, USEDFOR, Laying on}
Answer Source: KG
Answer: Laying on
Answer predicted: Loose (sic) thing in

Question 7: What can we do in this place shown in this image?
SPO triple:{Lake, USEDFOR, row a boat}
Answer Source: KG
Answer: row a boat
Answer predicted: build a treehouse

Question 8: This game is most popular in which country?
SPO triple:{Table Tennis, HASPROPERTY, Popular in China}
Answer Source: KG
Answer: Table Tennis
Answer predicted: Game Person Play

Question 9: What the baseball bat looks like?
SPO triple: {Baseball, ISA, Long round tapered object}
Answer Source: KG
Answer: long round tapered object
Answer predicted: aggressive animal

Figure 4.3: Performance of standalone Seeing is Knowing network without the composite score-based retrieval

be due to false positives in detections (see Questions 1-3 in Fig. 4.3). Giraffes, tomatoes, and foxes are mistaken to be camels, bell peppers, and lion respectively; entities relevant to the mistaken entities are then chosen as answers.

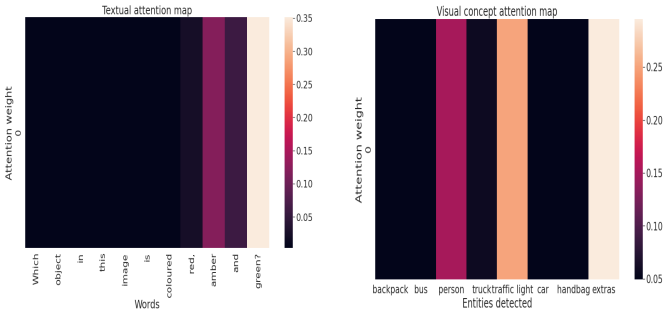
In cases where the language grounding is off, one case could be that the model has predicted the correct symbolism for a question (i.e. correct entity and relationship), but the retrieved entity comes from a fact that is not grounded in the language of the question. In Question 8 in Fig. 4.3, 'game person play', a property about the game table tennis, is chosen, but it is not grounded in the question-text which asks about the country where it is popular. A fix for this would be to ensure that entities' representations contain both lexical and graph-based features. One approach could be to use contextual word embeddings to encode entities before doing graph learning.

Lastly, entirely incorrect answers such as Question 9 in Fig. 4.3 can occur in case of sparse node connections (only one edge is present for 'aggressive animal'). SiK is indeed prone to making such unexplainable errors, highlighting the scope for improvement to be gained from incorporating better lexical semantic features as used in our composite score function.

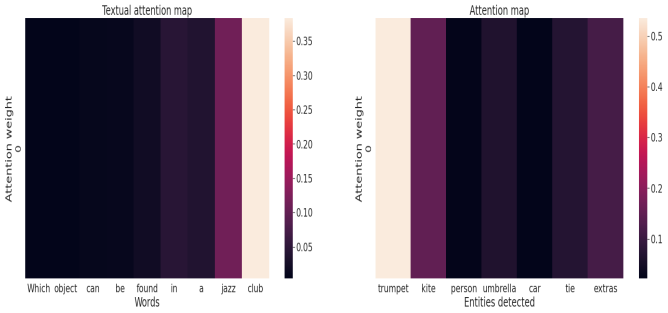
4.3.3 Visualization of the attention mechanism

To investigate the performance of the image-as-knowledge module, Fig. 4.4 depicts both the visual and textual attention maps learned by the network. We can see that the co-attention module is learning to pay attention to contextually relevant cues based on the input word embeddings in the question and the KG embeddings of the visual concepts. For Question 1 the textual attention correctly attends to the word 'amber and green' and the image attention correctly attends to the entity 'traffic light'. Similarly, in Question 2 the textual attention attends to the word 'jazz club' while the image attention attends to the entity 'trumpet'. Likewise in Question 3, words 'center of the image' and entity 'runway' are correctly attended to.

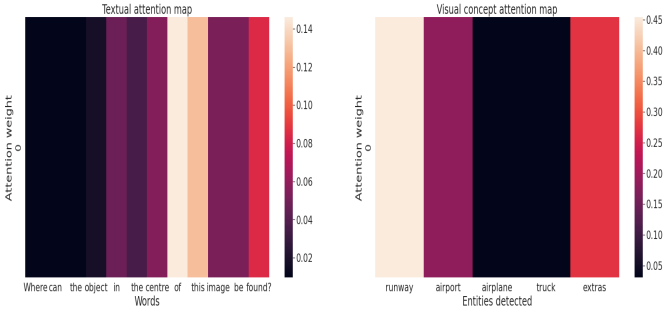
In cases where the answer is incorrectly predicted, the learned attention map sheds light on the reasoning of the network.



Question 1: Which object in this image is colored red, amber and green?
SPO triple: {Traffic Light, HasProperty, colour red amber and green}
Answer Source: Image
Answer: Traffic light
Answer predicted: **Traffic light**



Question 2: Which object can be found in a jazz club?
SPO triple: {Trumpet, AtLocation, jazz club}
Answer Source: Image
Answer: Trumpet
Answer predicted: **Trumpet**



Question 3: Where can object in the center of the image be found?
SPO triple: {Airplane, AtLocation, airport}
Answer Source: KG
Answer: Airport
Answer predicted: **Runway**

Figure 4.4: Visualizing co-attention maps produced by Seeing is Knowing

Table 4.6: Seeing is Knowing performance for different convex combinations of λ_k in the standard answer prediction task. λ_1 is the coefficient of KG similarity, λ_2 is the coefficient of Jaccard similarity and λ_3 is the coefficient for GLoVE similarity

Coefficient values			**Top 100 facts**		**Top 500 facts**	
λ_1	λ_2	λ_3	**Hits @1**	**Hits @3**	**Hits @1**	**Hits @3**
Single similarity metric being used						
1	0	0	56.05 ± 0.57	71.84 ± 0.44	55.43 ± 0.57	69.55 ± 0.58
0	1	0	17.44 ± 0.55	42.98 ± 0.86	13.83 ± 0.57	34.64 ± 1.00
0	0	1	13.38 ± 0.75	30.67 ± 0.80	11.75 ± 0.62	28.82 ± 0.75
Equal importance to all similarity metrics						
0.34	0.33	0.33	60.73 ± 0.82	77.03 ± 0.46	60.63 ± 0.73	76.9 ± 0.51
0.4	0.3	0.3	60.53 ± 0.94	$\mathbf{77.14 \pm 0.50}$	60.33 ± 0.92	77.01 ± 0.55
0.5	0.5	0	$\mathbf{60.82 \pm 0.98}$	75.83 ± 0.42	59.96 ± 0.89	74.48 ± 0.47
0.5	0.25	0.25	59.96 ± 0.72	76.73 ± 0.37	59.57 ± 0.68	76.31 ± 0.57
0.5	0	0.5	54.59 ± 71.76	71.76 ± 0.46	52.52 ± 0.74	71.03 ± 0.68
0	0.5	0.5	40.18 ± 1.13	61.62 ± 0.47	39.76 ± 1.08	60.84 ± 0.67
Highest importance to KG similarity						
0.6	0.2	0.2	58.89 ± 0.54	75.89 ± 0.44	58.5 ± 0.51	75.1 ± 0.658
0.7	0.15	0.15	58.05 ± 0.64	74.87 ± 0.49	57.6 ± 0.5	73.84 ± 0.49
0.8	0.1	0.1	57.23 ± 0.54	73.9 ± 0.56	56.73 ± 0.49	72.41 ± 0.40

4.4 Choosing hyperparameters for composite score-based retrieval

To investigate the importance of each metric in the composite score, we vary the respective coefficient λ_k to each metric across a range of values.

The first three rows of Table 4.6 demonstrate that the KG similarity metric individually is strongest by some margin. Considering only Jaccard similarity or GLoVe similarity is not enough to produce the accurate answer. This further highlights the crucial involvement that a KG reasoning module has in performance of any FVQA system. A standalone KG similarity metric, however, falls short of the highest achieved accuracy, owing to its relative inability to leverage lexical similarity between entity words and words in the question. (Note that the system obtained by assigning all the weight to only SiK metric is not the same as directly using SIK to retrieve the answer, since the former still uses the subgraph construction method to prune the fact search space. This is also demonstrated in the slightly different accuracies attained of the two methods.)

The results in Table 4.6 also demonstrate that lexical and distributional semantic features indeed hold complementary information to that of KG embeddings. To remedy the trend of SiK ignoring lexical similarity, roughly equal importance is

Table 4.7: Seeing is Knowing performance for different convex combinations of λ_k in the incomplete KG FVQA task for the top 100 facts retrieved. λ_1 is the coefficient of KG similarity, λ_2 is the coefficient of Jaccard similarity and λ_3 is the coefficient for GLoVE similarity

Coefficient values			**Only QA facts missing**		**50% KG missing**	
λ_1	λ_2	λ_3	**Hits @1**	**Hits @3**	**Hits @1**	**Hits @3**
Single similarity metric being used						
1	0	0	55.23 ± 0.44	70.56 ± 0.41	55.20 ± 0.77	70.47 ± 0.36
0	1	0	9.91 ± 0.58	27.44 ± 1.09	9.92 ± 0.61	27.48 ± 1.06
0	0	1	12.42 ± 0.72	29.43 ± 0.72	12.44 ± 0.76	29.47 ± 0.72
Equal importance to all similarity metrics						
0.34	0.33	0.33	55.13 ± 0.98	73.04 ± 0.54	55.13 ± 1.09	72.95 ± 0.38
0.4	0.3	0.3	56.35 ± 1.08	73.56 ± 0.58	56.38 ± 1.08	73.47 ± 0.44
0.5	0.5	0	54.78 ± 0.78	70.53 ± 0.56	54.79 ± 0.98	70.48 ± 0.33
0.5	0.25	0.25	$\mathbf{56.80 \pm 0.82}$	$\mathbf{73.62 \pm 0.42}$	$\mathbf{56.81 \pm 0.84}$	$\mathbf{73.54 \pm 0.25}$
0.5	0	0.5	53.68 ± 0.82	70.77 ± 0.67	53.63 ± 0.87	70.67 ± 0.35
0	0.5	0.5	26.8 ± 0.90	49.24 ± 0.62	26.79 ± 0.85	49.21 ± 0.63
Highest importance to KG similarity						
0.6	0.2	0.2	56.46 ± 0.51	73.36 ± 0.51	56.45 ± 0.63	73.31 ± 0.22
0.7	0.15	0.15	56.23 ± 0.36	72.7 ± 0.39	56.17 ± 0.66	72.63 ± 0.34
0.8	0.1	0.1	55.88 ± 0.25	72.01 ± 0.46	55.85 ± 0.65	71.91 ± 0.31

assigned to the Jaccard and GLoVE similarity metrics. Interestingly, the convex combination works best when each distance metric is weighted roughly equally. Our experiments indicate that the majority of the correct answers are produced due to the similarity score arising from SiK, but an equal weighting is required so as to nudge the correct answer ahead of the incorrect ones.

Lastly, we observe that weighting the SiK metric more than 0.6 starts to yield fewer benefits, and the performance monotonically decreases.

In both the incomplete KG settings (Table 4.7), we observe that pruning the fact search space still improves accuracy over the standalone networks. But the contribution of the textual similarity metrics becomes almost negligible when required facts are not present in the network.

4.5 Ethical impact

We now examine the ethical implications of our work. A prominent issue could be that of different biases known to exist in our data sources. Shankar et al. [60] showed population / representation bias existing in OpenImages and ImageNet.

Figure 4.5: **Question:** Where is this place?
SPO triple: : {Life, AtLocation, Zoo}
Answer Source: Knowledge Base
Answer: Zoo
Answer predicted: **in Zimbabwe**

Fisher et al. [61] showed web-scale common-sense KGs can be tough to curate and can allow biases to creep in. Janowicz et al. [62] note how the density of world locations generating DBPedia data (extracted from Wikipedia) is at odds with world population density. Fig. 4.5 shows how this manifests in our system. The answer for a place relevant to giraffes in wildlife, comes up as Zimbabwe, even though the only edges present in the KG concerning Zimbabwe were {Person, AtLocation, in Zimbabwe}, {Dog, AtLocation, in Zimbabwe}, {Tree, AtLocation, in Zimbabwe}, {Elephant, AtLocation, in Zimbabwe}. Such an error is quite informative and humbling – this means that different modalities (image, language, graph) working in tandem could still amplify their individual biases. For deploying such a system, we recommend debiasing of parameters learned in our architecture. While ConceptNet has been active in debiasing their representations, debiasing KG embeddings has not received as much attention, and it could pose subtle problems given that most entities would have low node-degree.

Chapter 5

End-to-end Fact-based Visual Spoken Question Answering

The previous chapters address the use of neuro-symbolic KG embeddings for reasoning over incomplete KGs. One obvious use-case for incorporating incomplete external knowledge is voice-assistants, which are ubiquitous across households in today's age. Although such voice assistants have started providing services across many different languages, there still exists an asymmetry in the research community in terms of attention paid to the study of such benchmarks across these different languages. Well-resourced languages generally also have mature automatic speech recognition (ASR) systems and language models. The accompanying KGs also tend to be limited to languages that are well-resourced [5, 6, 7]. Against this background, it is worthwhile to think of building end-to-end systems which directly use speech signals as input which can readily harness huge knowledge repositories stored in another language, instead of requiring tabula rasa learning. The key idea is this: Since entities and concepts in a KG are symbolic, they should be language-agnostic. Can they be readily transferred to under-resourced languages?

With these motivations, the main contributions of our next work Worldly-Wise [8] are two-fold: 1) A new task referred to as Fact-based Visual Spoken-Question Answering (FVSQA) along with the release of 5 hours of synthetic-speech data in each of the three languages - English, Hindi, and Turkish. 2) An end-to-end architecture Worldly-Wise (WoW) capable of answering questions trained directly on speech features in all three languages. Fig. 5.1 shows an illustrative example of the type of questions in the dataset. To the best of our knowledge, this is the first work to perform KG knowledge acquisition using only a speech signal as input, without the requirement for a pre-trained automatic speech recognizer as a system component.

Worldly-Wise (WoW) is readily generalizable to other languages, even those without an ASR-system. This is possible for two reasons - a) it obtains speech features as mel-frequency cepstral coefficients (MFCCs) and does not require ASR-based text-conversion or speech feature extraction from a language-specific

Figure 5.1: Example of a fact-based visual question
Question - Which object in this image can be found in a jazz club?
Supporting fact - You are likely to find [[a trumpet]] in [[a jazz club]]
Subject, Predicate, Object - (Trumpet, AtLocation, Jazz Club)
Answer - Trumpet

pretrained network, and b) for knowledge acquisition, it does not require the entity label to be in the same language as the question, instead leveraging neuro-symbolic entity representations in the form of KG embeddings. These KG embedding methods, trained to remedy KG sparsity by performing missing-edge prediction, learn transferable entity-features that encode the local and global structures in KGs. Worldly-wise uses the same image-as-knowledge representation as Seeing is Knowing. But in this case, it is applied to a speech signal as opposed to a textual question.

We report experimental results on synthetic speech data in the aforementioned diverse languages to demonstrate its effectiveness. Hindi and Turkish are simulated as under-resourced languages by denying the system access to any text, ASR, or machine translation to or from those languages, thereby requiring the system to learn the mapping from Hindi and Turkish speech signals to the KG knowledge stored in English. Through this work, we hope to motivate research in expanding spoken language understanding (SLU) in under-resourced languages through models which circumvent the need for parallel text-labelled resources.

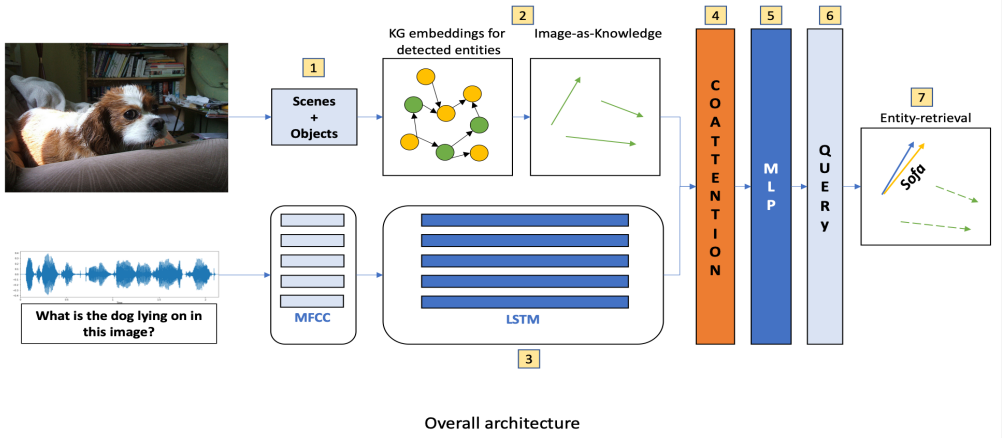


Figure 5.2: Our architecture for FVSQA. (1) Scenes & objects are detected in the image. (2) For detected entities, we retrieve their KG embedding representations. The span of these embedding vectors represents the image-as-knowledge. (3) MFCC features for the spoken question are accumulated via an LSTM. (4) The joint image-question encoding is derived using an attention mechanism described in Fig. 5.3, then (5) passed through a multi-layer perceptron, whose (6) last layer is used as a query that is (7) used to retrieve the entity to answer the question.

5.1 Related work: Multimodal SLU

Spoken language understanding (SLU) has a long history. For most of its history, SLU was developed in a pipelined fashion, with ASR feeding text to a natural language understanding system; for example, to the best of our knowledge, the only published uses of SLU with knowledge graphs that fits this description is [63]. Recent research in end-to-end multimodal SLU bypasses the need for ASR by leveraging a parallel modality such as image [64, 65] or video [66], or a non-parallel corpus of text [67], to guide learning speech embeddings such that the speech input can be used in a downstream task.

In speech-based VQA applications, the most common approach is a two-step approach which consists of an ASR followed by text-based VQA [68]. However, these systems are not generalizable to under-resourced or unwritten languages for which we cannot train an ASR system. Therefore, in this study, we will explore using neural speech embeddings, which are guided by the information in the KG, for achieving FVSQA.

5.2 Task formulation

This section introduces a new task called FVSQA and presents a new dataset collected for this task.

5.2.1 FV(S)QA

FVSQA is identical to FVQA, the difference being the modality of the question q; in FVSQA is a speech signal instead of a textual question. The following condition holds for questions in the (FVQA [1]) benchmark: For each (question,image,answer) triplet in the dataset ($(q_i, I_i, y_i) \in \mathcal{D}$), exactly one supporting fact in the knowledge graph ($f_j = (h, r, t) \in \mathcal{G}$) exists such that the correct answer y_i is either the head or the tail of f_j, and such that at least one of the two entities is visible in the image.

The companion KG for the dataset is constructed from three diverse sources: Webchild [6], ConceptNet [7], and DBPedia [5]. DBPedia provides parent-child relationships between different entities, ConceptNet provides common-sense knowledge about entities, whereas Webchild provides many different kinds of comparative relationships between entities (comparative relations are considered as a single relationship type for FVQA).

Answering questions in FVQA is to perform the following operation:

$$\hat{y} = \underset{e \in \mathcal{E}}{\operatorname{argmax}}\ p(y = e \mid q, I, \mathcal{G}), \tag{5.1}$$

i.e., retrieving the most probable entity as the answer given a question q and image I, and given the graph $\mathcal{G}$.

The FVSQA task formulation is identical, except that the question is not textual but spoken. We study the task when the question is spoken in one of three languages: English, Hindi, Turkish.

5.2.2 Data description

The dataset consists of 2190 images sampled from the ILSVRC [39] and the MSCOCO [40] datasets. There are 5826 questions concerning 4216 unique facts (Table 2.1). FVSQA provides the same five train-test splits as FVQA, where each split contains images and questions in the ratio 1:1. The accompanying KG consists

of roughly 194500 facts, concerning 88606 entities. In total, the dataset contains 13 relations: $R \in \{$*Category, HasProperty, RelatedTo, AtLocation, IsA, HasA, CapableOf, UsedFor, Desires, PartOf, ReceivesAction, CreatedBy, Comparative*$\}$.

The multi-lingual speech data generation procedure is described next.

Data Generation - Text Translation

The text questions in FVSQA dataset are in English. To generate spoken questions in Hindi and Turkish, we first translate the questions using Amazon Translate API[1] from English. We manually review the questions to ensure intelligibility of questions. These translated texts are only used for speech data generation; these are not available to the network during either training or inference.

Data Generation - Text-to-Speech

We use Amazon's Polly API[2] to generate spoken questions for each language. The generated speech is in mp3 format, sampled at 22 kHz. For a given language, all questions were generated using the same voice. The voices used were Joanna for English, Aditi for Hindi, and Filiz for Turkish. We again manually review and ensure intelligibility of speech data so generated.

5.3 Our approach

The proposed architecture for FVSQA is shown in Fig. 5.2. As shown in the figure, attention fuses an image I and question q to form a query vector ν. This query vector is then used to retrieve the answer from the KG as

$$\hat{y}(q|I) = \underset{e \in \mathcal{E}}{\operatorname{argmax}}\, \nu(q, I)^T e. \tag{5.2}$$

The image and KG representations (depicted again in Figures 5.2 and 5.3) are similar to those considered in the previous chapter, but the question representation is different (described below). Furthermore, the goal of SiK is different from WoW, as the former performs monolingual text-FVQA over incomplete KGs.

[1] https://aws.amazon.com/translate/
[2] https://aws.amazon.com/polly/

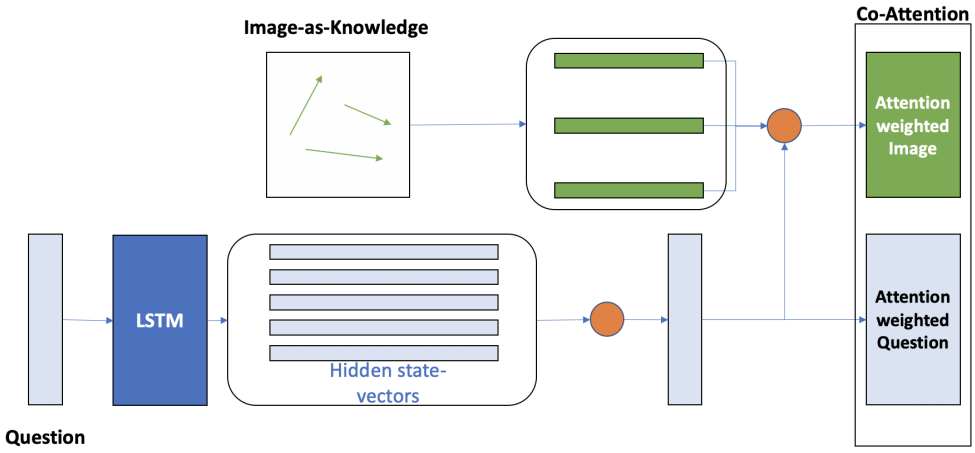


Figure 5.3: Image and query are fused using the co-attention mechanism depicted here. First, self-attention computes a weighted summary of the speech signal (bottom orange circle). Second, the summary is used to compute an attention-weighted summary of the concepts in the image (top orange circle). The resulting image query is a vector drawn from the span of the entities present in the image.

Table 5.1: KG embedding accuracy

MR	MRR	Hits@1	Hits@3	Hits@10
11194	0.156	0.132	0.152	0.197

5.3.1 Question representation

We represent the speech waveforms using MFCC features. We set the window length to 25 ms and stride size of 10 ms. For each time-step, we follow standard convention of using 39-dimensional vectors: the first 12 cepstral coefficients and the energy term, along with delta and double-delta features to gather contextual information.

Table 5.2: FVSQA performance of WoW architecture across different languages

Language	Hits @1	Hits @3
English	49 ± 0.62	61.85 ± 1.13
Turkish	48.96 ± 1.14	61.56 ± 0.79
Hindi	49.29 ± 0.73	61.26 ± 0.93
English - ASR + text-FVQA	54.07 ± 1.15	65.52 ± 0.75

5.4 Results and discussion

The experimental setup is similar to that described in the previous chapter.

5.4.1 Cross-lingual FVSQA

FVSQA is trained using the best performing KG embedding model demonstrated in [2] and its performance is highlighted in Table 5.1. To verify the superiority of ERMLP over word-embeddings, we compare a model trained with KG entities represented as averaged word embeddings instead. This representation fails to train an end-to-end system even for English, the final accuracy being close to 0%.

Aided by ERMLP, WoW is able to perform FVSQA at the same levels of accuracy across English, Hindi, and Turkish (see Table 5.2). For English, we additionally investigate an ASR + text-based system, where the FVQA model is trained on gold-standard textual questions, and during inference-time, an ASR-converted speech transcript of the question is provided. The ASR system is based on the pre-trained Kaldi ASpIRE model[3] which was originally trained on the augmented Fisher English dataset. The resulting FVQA system performs better than an end-to-end system for English, indicating that some joint-training strategies for speech and text-based systems could help increase accuracy for the end-to-end speech system. However, our experiments on sharing the lower layers of the network between speech and text-systems did not improve accuracy of the end-to-end speech system for English.

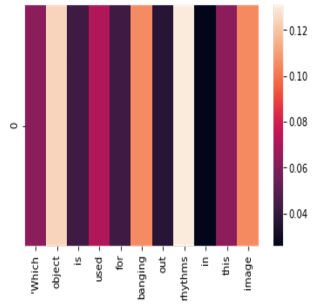
5.4.2 Attention mechanism visualizations

We can see in Fig. 5.4 that for each language, the speech signal can perform as a good query vector to calculate contextual visual attention as per Eq. (3.6). The resulting IaK attention maps are interpretable and, in cases where the network predicts the wrong answer, provide an insight into the reason for the network's failure (Fig. 5.4).

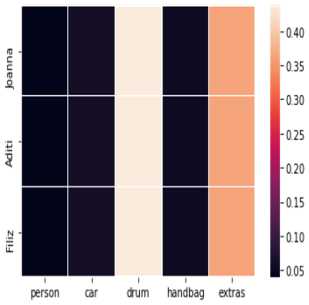
Furthermore, the speech self-attention maps are also coherent and informative. The questions in Fig. 5.4 show attention accumulated by each word over the time-steps of the word's utterance. We can clearly see that the relevant time-

[3] https://kaldi-asr.org/models/m1

Question 1: Which object is used for banging out rhythms in this image?
SPO triple: {Drum, UsedFor, banging out rhythms}
Answer Source: Image
Answer: Drum
Answer predicted: **Drum**

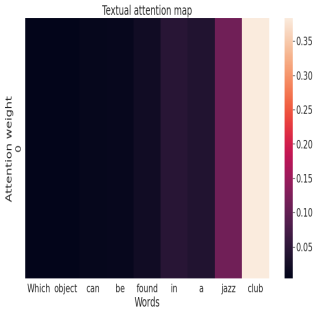


Speech attention summed over all time-steps of a word

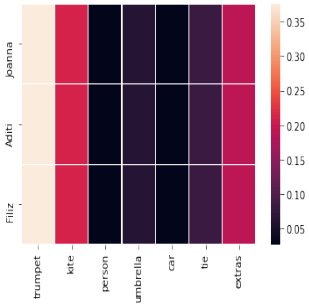


Visual Attention for all three voices

Question 2: Which object in this image can be found in a jazz club?
SPO triple: {Trumpet, AtLocation, jazz club}
Answer Source: Image
Answer: Trumpet
Answer predicted: **Trumpet**

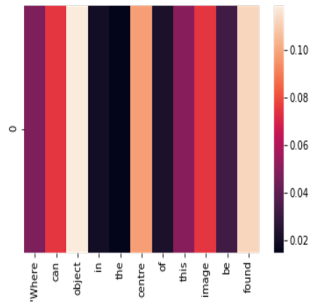


Speech attention summed over all time-steps of a word

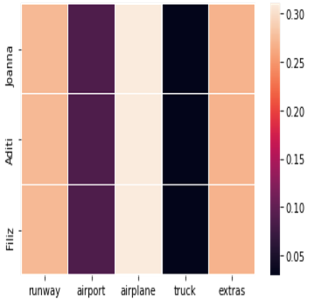


Visual Attention for all three voices

Question 2: Where can object in the center of image be found?
SPO triple: {Airplane, AtLocation, Airport}
Answer Source: Image
Answer: Airport
Answer predicted: **Runway**



Speech attention summed over all time-steps of a word



Visual Attention for all three voices

Figure 5.4: Visualizing co-attention maps produced by Worldly-Wise

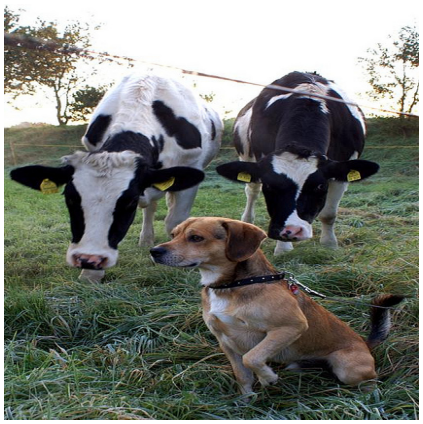


Figure 5.5: **Question** - Which animal in this image is man's best friend?
Supporting fact - [[dogs]] are [[man's best friend]]
Subject, Predicate, Object - (Dog, HasProperty, man's best friend)
Answer - Dog

steps are attended to, depending on the image and the question itself. To the best of our knowledge, this is the first work to jointly learn attention-based speech representations guided by external KG knowledge.

5.5 Ethical impact

We now turn to the ethical implications of this work. Worldly-Wise relies on leveraging cross-lingual knowledge resources for question answering. While this approach yields enormous benefits, care must be taken to evaluate appropriateness of the source of knowledge depending on the language. What may be considered as conventional wisdom in one culture or language may not be so for another. An example of how this manifests in our dataset is shown in Fig. 5.5. The knowledge graph conveys conventional wisdom in English that 'A dog is man's best friend', and therefore the expected answer to this question is 'Dog'. However, in regions where Hindi is spoken, the answer could equally be expected to be 'Cow' that appears in the image. This example is quite informative, and if such an instance can occur in the extreme, it could lead to fairness issues. This highlights the fundamental tradeoff involved in training such a cross-lingual system on knowledge generated in another language. Governance of such a system is therefore essential to ensure cultural appropriateness and fairness in different contexts.

Chapter 6

Conclusion and Future Work

6.1 Concluding remarks

Motivated by the seamless nature of human perception and reasoning, this thesis discussed two works which try to impart relational semantics fluidly across visual, textual, and audio inputs (that too in different languages). Hopefully, such mapping across domains can help machines execute reasoning tasks in a robust, explainable, and generalizable fashion.

The first work, Seeing is Knowing, solves the FVQA task even when the required edge is missing from the knowledge graph. In the process, we present the first approach to use KG embeddings for VQA in general. A composite answer retrieval method augments its accuracy by incorporating complementary lexical features and KG semantic features. Serendipitous benefits of the standalone approach include: (1) improved computational complexity, and (2) an improved representation of each image, as the span of KG embeddings of the visible entities. Future work might consider methods that combine global features via KG embeddings and local features via GCN for this task, in order to strengthen the KG-based entity embedding.

The second work introduced a new task, FVSQA, along with an architecture that can perform cross-lingual knowledge acquisition for question-answering. In the process, we demonstrate the first task to perform knowledge acquisition directly using a speech signal as an input. This knowledge acquisition for speech can be extended to other tasks such as audio caption-based scene identification [64] and multi-modal word discovery [69]. Future work will include extending FVSQA to a multi-speaker setting, gathering spoken data from real-world speakers, as well as extending it to languages without an ASR system.

6.2 A vision for the future

A famous Sanskrit couplet by the poet *Kalidasa* begins with the phrase *Vagartha-viva Sampruktau* - roughly translating as 'sound and meaning are inextricably linked'. In the previous chapters, we discussed architectures which transform input text, speech, images, and knowledge graphs and process them together in order to perform the semantic task of Fact-based Visual Question Answering. A natural question to ask is: What if the knowledge graph itself is spoken?

6.2.1 Spoken knowledge graphs - From sound to meaning

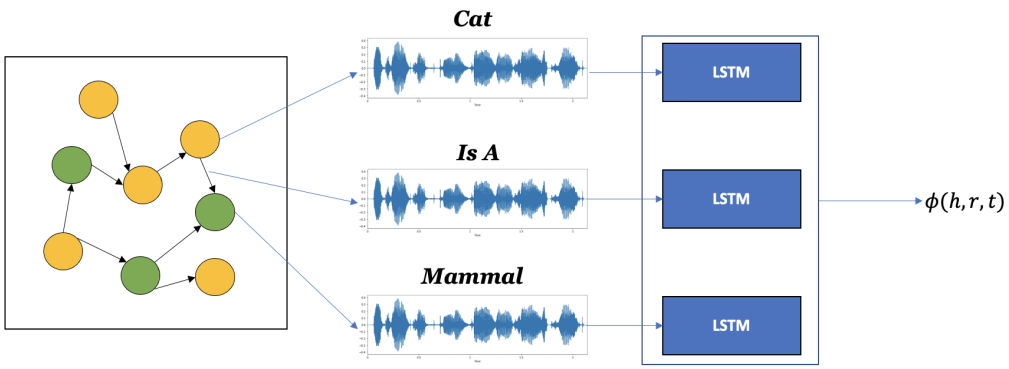


Figure 6.1: Concept design for spoken knowledge graphs

Indeed, the idea of multimodal knowledge bases has been explored before [70]. However, knowledge graphs where the entities are spoken waveforms and thus have a representation in the time-frequency domain are yet to be explored. Babies learn to recognize words and concepts from audio inputs long before they can read or write [64]. Their experiences with the outside world and multi-modal sensory inputs help them learn associations between concepts, and the various modalities in which these concepts can manifest themselves.

From a scientific perspective, it would be interesting to analyze the emergent salient features in audio representations due to graph-based connections. Figure 6.1 describes what such a system might look like. By processing the speech waveforms in an end-to-end fashion, one could hypothesize that phoneme / morpheme / word-like units could arise from raw speech inputs owing to the semantic structures imposed by the KG.

Such a spoken knowledge graph can have various practical benefits too. The learned LSTMs can be used as pre-training for various semantic and multimodal tasks involving speech inputs such as speech recognition, spoken question answering, audio transcription, and audio translation.

References

[1] P. Wang, Q. Wu, C. Shen, A. van den Hengel, and A. R. Dick, "FVQA: Fact-based visual question answering," *CoRR*, vol. abs/1606.05433, 2016. [Online]. Available: http://arxiv.org/abs/1606.05433

[2] K. Ramnath and M. Hasegawa-Johnson, "Seeing is Knowing! Fact-based visual question answering using knowledge graph embeddings," *ArXiv*, vol. abs/2012.15484, 2020.

[3] H. Li, P. Wang, C. Shen, and A. V. D. Hengel, "Visual question answering as reading comprehension," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6312–6321, 2019.

[4] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, and Z. Ives, "DBPedia: a nucleus for a web of open data," in *PROC. 6TH INT'L SEMANTIC WEB CONF*. Springer, 2007.

[6] N. Tandon, G. de Melo, F. M. Suchanek, and G. Weikum, "Webchild: Harvesting and organizing commonsense knowledge from the web," in *Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, NY, USA, February 24-28, 2014*, 2014, pp. 523–532.

[7] H. Liu and P. Singh, "ConceptNet: a practical commonsense reasoning toolkit," *BT Technology Journal*, vol. 22, pp. 211–226, 2004.

[8] K. Ramnath, L. Sari, M. Hasegawa-Johnson, and C. D. Yoo, "Worldly wise! Cross-lingual knowledge fusion for fact-based visual spoken-question answering," June 2021.

[9] R. F. Simmons, *Synthetic Language Behavior*. System Development Corporation, 1963.

[10] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A core of semantic knowledge," in *PROC. OF WWW '07*. ACM, 2007, pp. 697–706.

[11] T. Mitchell and E. Fredkin, "Never-ending language learning," in *2014 IEEE International Conference on Big Data (Big Data)*, 2014, pp. 1–1.

[12] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *SIGMOD Conference*, 2008, pp. 1247–1250.

[13] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. P. Murphy, T. Strohmann, S. Sun, and W. Zhang, "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," in *Knowledge Distillation and Data Mining*, 2014, pp. 601–610.

[14] G. S. Halford, W. H. Wilson, and S. Phillips, "Relational knowledge: The foundation of higher cognition," *Trends in Cognitive Sciences*, vol. 14, pp. 497–505, 2010.

[15] G. Murphy, *The Big Book of Concepts*. MIT press, 2004.

[16] A. Bordes, N. Usunier, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in Neural Information Processing Systems 26. Curran Associates, Inc*, 2013, pp. 2787–2795.

[17] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang, "Rotate: Knowledge graph embedding by relational rotation in complex space," in *Proc. International Conference on Learning Representations*, 2019, pp. 1–18.

[18] R. Socher, D. Chen, C. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledgebase completion," in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2013, p. 926–934.

[19] M. Nickel, V. Tresp, and H. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proc. International Conference on Machine Learning*, 2011.

[20] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2D knowledge graph embeddings," in *Proc. AAAI*, S. Zilberstein, S. McIlraith, and K. Weinberger, Eds., 2018, pp. 1811–1818.

[21] A. Bordes, N. Usunier, S. Chopra, and J. Weston, "Large-scale simple question answering with memory networks." *CoRR*, vol. abs/1506.02075, 2015.

[22] D. Lukovnikov, A. Fischer, J. Lehmann, and S. Auer, "Neural network-based question answering over knowledge graphs on word and character level," in *Proceedings of the 26th International Conference on World Wide Web*, ser. WWW '17. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017. [Online]. Available: https://doi.org/10.1145/3038912.3052675 p. 1211–1220.

[23] X. Huang, J. Zhang, D. Li, and P. Li, "Knowledge graph embedding based question answering," in *ACM International Conference on Web Search and Data Mining*, 2019.

[24] Z. Dai, L. Li, and W. Xu, "CFO: Conditional focused neural question answering with large-scale knowledge bases," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Aug. 2016, pp. 800–810.

[25] A. Saxena, A. Tripathi, and P. Talukdar, "Improving multi-hop question answering over knowledge graphs using knowledge base embeddings," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.412 pp. 4498–4507.

[26] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on Freebase from question-answer pairs," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, Oct. 2013. [Online]. Available: https://www.aclweb.org/anthology/D13-1160 pp. 1533–1544.

[27] Y. Zhang, H. Dai, Z. Kozareva, A. J. Smola, and L. Song, "Variational reasoning for question answering with knowledge graph," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[28] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "OK-VQA: A visual question answering benchmark requiring external knowledge," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[29] N. Y. Sanket Shah, Anand Mishra and P. P. Talukdar, "KVQA: Knowledge-aware visual question answering," in *Proc. AAAI Conference on Artificial Intelligence*, 2019.

[30] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2016. [Online]. Available: https://arxiv.org/abs/1602.07332

[31] J. Lei, L. Yu, T. Berg, and M. Bansal, "TVQA+: Spatio-temporal grounding for video question answering," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.730 pp. 8211–8225.

[32] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 1682–1690.

[33] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2425–2433.

[34] Y. Zhu, O. Groth, M. S. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4995–5004, 2016.

[35] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *NIPS*, 2015.

[36] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *CVPR*, 2017.

[37] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "MovieQA: Understanding stories in movies through question-answering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[38] C. Fellbaum and G. Miller, Eds., *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[39] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, 2015.

[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick, "Microsoft COCO: Common objects in context," *Lecture Notes in Computer Science*, vol. 8693, pp. 740–755, 2014.

[41] M. Narasimhan and A. G. Schwing, "Straight to the facts: Learning knowledge base retrieval for factual visual question answering," *CoRR*, 2018.

[42] M. Narasimhan, S. Lazebnik, and A. Schwing, "Out of the box: Reasoning with graph convolution nets for factual visual question answering," in *Advances in Neural Information Processing Systems 31*, 2018, pp. 2654–2665.

[43] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[44] L. Cai and W. Y. Wang, "KBGAN: Adversarial learning for knowledge graph embeddings," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, June 2018.

[45] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," vol. arXiv:1406.2661, 2014.

[46] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *ArXiv*, vol. abs/1701.07875, 2017.

[47] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," ser. Proceedings of Machine Learning Research, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010, pp. 297–304.

[48] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for self-supervised learning of language representations," in *International Conference on Learning Representations*, 2020.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image classification," in *Proc. International Conference on Learning Representations*, 2015.

[51] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 91–99.

[52] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[53] ZFTurbo, "Keras-RetinaNet for open images challenge 2018," https://github.com/ZFTurbo/Keras-RetinaNet-for-Open-Images-Challenge-2018.git, 2018.

[54] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Van Den Hengel, "What value do explicit high level concepts have in vision to language problems?" in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 203–212.

[55] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *BMVC*, 2016.

[56] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[57] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Oct. 2014, pp. 1532–1543.

[58] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances in Neural Information Processing Systems 29*, 2016, pp. 289–297.

[59] K. Guu, J. Miller, and P. Liang, "Traversing knowledge graphs in vector space," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Sep. 2015, pp. 318–327.

[60] S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. Wilson, and D. Sculley, "No classification without representation: Assessing geodiversity issues in open data sets for the developing world," in *NIPS 2017 workshop: Machine Learning for the Developing World*, 2017.

[61] J. Fisher, D. Palfrey, C. Christodoulopoulos, and A. Mittal, "Measuring social bias in knowledge graph embeddings," vol. arXiv preprint arXiv:1912.02761, 2019.

[62] K. Janowicz, B. Yan, B. Regalia, R. Zhu, and G. Mai, "Debiasing knowledge graphs: Why female presidents are not like female popes," in *Proceedings of the ISWC 2018 Posters & Demonstrations*, 2018.

[63] W. A. Woods, "Motivation and overview of SPEECHLIS: An experimental prototype for speech understanding research," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 2–10, 1975.

[64] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *Advances in Neural Information Processing Systems*, 2016, pp. 1858–1866.

[65] H. Kamper, A. Anastassiou, and K. Livescu, "Semantic query-by-example speech search using visual grounding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7120–7124.

[66] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: A large-scale dataset for multimodal language understanding," *arXiv preprint arXiv:1811.00347*, 2018.

[67] L. Sarı, S. Thomas, and M. Hasegawa-Johnson, "Training spoken language understanding systems with non-parallel speech and text," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 8109–8113.

[68] T. Zhang, D. Dai, T. Tuytelaars, M.-F. Moens, and L. Van Gool, "Speech-based visual question answering," *arXiv preprint arXiv:1705.00464*, 2017.

[69] D. Harwath, G. Chuang, and J. Glass, "Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, 2018, pp. 4969–4973.

[70] P. Pezeshkpour, L. Chen, and S. Singh, "Embedding multimodal relational data for knowledge base completion," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018. [Online]. Available: https://www.aclweb.org/anthology/D18-1359 pp. 3208–3218.